Journal of Mathematics and Physics

(Founded by C. L. E. Moore)

Volume XXXV

1956

Board of Editors

E. REISSNER, *Managing Editor*

H. FESHBACH
P. FRANKLIN
E. A. GUILLEMIN
C. C. LIN
P. M. MORSE
D. J. STRUIK
N. WIENER

The Technology Press
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

QA 1
J87
v.35,

CONTENTS

Page

The Flow of a Viscous Liquid past a Flat Plate at Small Reynolds Numbers
By S. Tomotika and H. Yosinobu 1

Small Perturbations in the Unsteady Flow of a Compressible Viscous and Heat-Conducting Fluid
By T. Yao-tsu Wu 13

Vibration and Stability Problems of Beams Treated by Matrices
By K. Marguerre 28

Stress Theory of Plastic Flow
By D. Trifan 44

On the Problem of Buckling of a Hyperbolic Paraboloidal Shell Loaded by its own Weight
By Anthony Ralston 53

Analysis of an Infinite Plate Containing Radial Cracks Originating at the Boundary of an Internal Circular Hole
By O. L. Bowie 60

Formulae Relating Some Equivalent Networks
By R. J. Duffin and Elsa Keitzer 72

An Existence Theorem for Driving-Point Impedance Functions
By N. Declaris 83

On Discrete Operators Connected with the Dirichlet Problem
By Mervin E. Muller 89

Inverse Laplace Transforms Expressed as Neumann Series
By Enzo Cambi 114

A Method of Solving Inhomogeneous Linear Simultaneous Equations
By F. B. Stulen and F. G. Lehman 123

Approximation and Projection
By Arthur Sard 127

The Solution of the Diffusion Equation by a High Order Correct Difference Equation
By Jim Douglas, Jr. 145

Coefficients for Complex Osculatory Interpolation over a Cartesian Grid
By Herbert E. Salzer 152

A Power Series Solution of the Nonlinear Equations for Axi-Symmetrical Bending of Shallow Spherical Shells
By Roger M. Simons 164

Some Definite Integrals Involving Conical Functions
By Leopold B. Felsen 177

Compressible Viscous Flow Past a Wedge Moving at Hypersonic Speeds
By L. J. Pan and Y. H. Kuo 179

Rayleigh's Problem at Low Mach Number According to the Kinetic Theory of Gases
By Hsun Tiao Yang and Lester Lees 195

A Free Streamline Theory for Two-Dimensional Fully Cavitated Hydrofoils
By T. Yao-tsu Wu 236

The Magnetostatic Characteristics of Two Non-Magnetic Elliptic Cylinders
By E. E. Jones 266

Analysis of a General System for the Detection of Amplitude-Modulated Noise
By Emanual Parzen and Norman Shiren 278

On the Ratio of Consecutive Eigenvalues
By L. E. Payne, G. Polya and H. F. Weinberger 289

The Truncation Error in a Semi-Discrete Analog of the Heat Equation
By Norman E. Friedmann 299

Page

Note on the Sensitivity of Least Squares Solutions
By Mark Lotkin 309
A Vector Method for Solving Linear Equations and Inverting Matrices
By J. P. Roth and D. S. Scott 312
A Note on Salzer's Method for Summing Certain Slowly Convergent Series
By P. Wynn 318
A Note on the Solution of a Non-Linear Transonic Flow Equation
By John T. Robacker and Albert D. Wheelon 321
Elastic Waves in Anisotropic Media
By J. L. Synge 323
Deformation Energy and the Stress-Strain Relations for Isotropic Materials
By T. Y. Thomas 335
A Derivation of the Equations of Shell Theory for General Orthogonal Coordinates
By James K. Knowles and Eric Reissner 351
On the Unsteady Motion of a Viscous Fluid Past a Semi-Infinite Flat Plate
By G. F. Carrier and R. C. DiPrima 359
One-dimensional Shock-Motion of a Gas With $\gamma = 3$, Enclosed Between Fixed Walls
By G. Kuerti and K. Faymon 384
Osculatory Interpolation and Integration
By J. C. Cooke 394
Limits of the Characteristic Values for Certain Boundary-Value Problems Associated with Difference Equations
By Tomlinson Fort 401
Improved Approximations to Differential Equations by Difference Equations
By W. Robert Mann, C. L. Bradshaw and J. Grady Cox 408

THE FLOW OF A VISCOUS LIQUID PAST A FLAT PLATE AT SMALL REYNOLDS NUMBERS

By S. Tomotika and H. Yosinobu

1. Introduction. A second approximation for the skin friction at small Reynolds numbers of a flat plate placed edgewise along a steady uniform stream has been calculated by Piercy and Winny [1], Davies [2], and Sidrak [3] respectively, by developing different analyses from each other. But, as mentioned in detail in our previous paper [4], no complete agreement can be found among these three expansion formulae, and they differ slightly from each other.

Recently we have examined very carefully the work by these writers, and it has been found that Piercy and Winny's formula is correct, but the remaining two are both erroneous.

Davies attributes the discrepancy between his own formula and Piercy and Winny's to the fact that Southwell and Squire's equations of motion, upon which his analysis is based, are different from Oseen's equations upon which Piercy and Winny's analysis is based. It is readily found however that in a particular case of a flat plate placed edgewise along a uniform stream, Southwell and Squire's equations are quite equivalent to Oseen's equations. Therefore it is expected that an expansion formula in complete agreement with Piercy and Winny's may be obtained by pursuing Davies's analysis, and to this enquiry the present paper is directed.

On the other hand, Aoi and the senior writer [5] have recently carried out a detailed study on the slow motion of an incompressible viscous fluid past an elliptic cylinder on the basis of Oseen's equations, and a general expression as well as an expansion formula correct to the fourth power of the Reynolds number have been obtained for the drag of the elliptic cylinder placed parallel or perpendicularly to the direction of the undisturbed uniform stream. Also, a third approximation for the skin friction of a flat plate has been derived as a limiting form of the expansion formula for the drag of the elliptic cylinder, and it has thus been found that up to the second approximation, our formula is in perfect agreement with Piercy and Winny's.

Further, the present writers [6] have succeeded to derive the correct second approximation for the skin friction of the flat plate by developing correct analysis along the lines of Sidrak's work.

2. The extended Oseen's equations of motion and their fundamental solutions. According to Lewis [7], the extended Oseen's equations of motion due originally to Southwell and Squire for the two-dimensional steady flow of an incompressible viscous fluid can be written in the forms:

$$\nu \frac{\partial \zeta}{\partial \beta} = -\frac{\partial P}{\partial \alpha}, \qquad -\zeta - \nu \frac{\partial \zeta}{\partial \alpha} = -\frac{\partial P}{\partial \beta}, \tag{2.1}$$

where the vorticity ζ and the pressure function P are defined by the relations

$$\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \qquad P = \frac{p}{\rho} + \tfrac{1}{2}(u^2 + v^2), \tag{2.2}$$

and u, v are the Cartesian components of velocity, p the pressure, and ρ and ν the density and the kinematic coefficient of viscosity of the fluid respectively. α and β are respectively the velocity potential and the stream function for the flow of an incompressible perfect fluid past a body concerned. Thus, in a particular case when a steady uniform stream of velocity U in the positive direction of the x-axis flows past a flat plate lying on the x-axis, we have

$$\alpha + i\beta = -U(x + iy). \tag{2.3}$$

It is easily verified from equations (2.1) that both ζ and P satisfy the partial differential equation:

$$\frac{\partial \Phi}{\partial \alpha} + \nu\left(\frac{\partial^2 \Phi}{\partial \alpha^2} + \frac{\partial^2 \Phi}{\partial \beta^2}\right) = 0, \tag{2.4}$$

and if one of them is known the other can be found by quadrature; if, for example, ζ is known, P can be found as:

$$P = \int \left\{ -\nu \frac{\partial \zeta}{\partial \beta}\, d\alpha + \left(\zeta + \nu \frac{\partial \zeta}{\partial \alpha}\right) d\beta \right\}. \tag{2.5}$$

Alternatively, both can be derived from a single function Φ satisfying equation (2.4) by means of the relations:

$$\zeta = \frac{\partial \Phi}{\partial \alpha}, \qquad P = -\nu \frac{\partial \Phi}{\partial \beta}. \tag{2.6}$$

In order to solve equation (2.4) we put

$$\Phi = \Phi' e^{-\alpha/2\nu}, \tag{2.7}$$

so that Φ' satisfies the equation:

$$\frac{\partial^2 \Phi'}{\partial \alpha^2} + \frac{\partial^2 \Phi'}{\partial \beta^2} = \frac{1}{4\nu^2}\Phi'. \tag{2.8}$$

Now we introduce the elliptic coordinates (ξ, η) connected with the Cartesian coordinates as:

$$x + iy = c \cosh(\xi + i\eta). \tag{2.9}$$

Then, a flat plate of length $2c$ lying on the x-axis, with its mid-point at the origin of the coordinate-axes, is expressed by

$$\xi = 0. \tag{2.10}$$

Equation (2.3) now becomes

$$\alpha + i\beta = -Uc \cosh(\xi + i\eta), \tag{2.11}$$

and if, in place of α, β, we make use of ξ, η as independent variables, we find that Φ' satisfies the equation:

$$\frac{\partial^2 \Phi'}{\partial \xi^2} + \frac{\partial^2 \Phi'}{\partial \eta^2} = k^2(\cosh^2 \xi - \cos^2 \eta)\Phi', \tag{2.12}$$

where

$$k = Uc/(2\nu). \tag{2.13}$$

Solving this equation by substituting $\Phi' = F(\xi)G(\eta)$ and separating the variables, we are led to the modified Mathieu equation for $F(\xi)$ and the Mathieu equation for $G(\eta)$, namely

$$\frac{d^2F}{d\xi^2} - (\lambda + k^2\cosh^2\xi)F = 0, \tag{2.14}$$

$$\frac{d^2G}{d\eta^2} + (\lambda + k^2\cos^2\eta)G = 0. \tag{2.15}$$

Solutions of equation (2.15) with period 2π with respect to η are obtained for a discrete set of characteristic values of λ. They fall into the four groups. Thus,

$$\left.\begin{aligned} ce_{2n}(\eta) &= \sum_{r=0}^{\infty}(-1)^{n+r}A_{2r}^{(2n)}\cos 2r\eta, \\ ce_{2n+1}(\eta) &= \sum_{r=0}^{\infty}(-1)^{n+r}B_{2r+1}^{(2n+1)}\cos(2r+1)\eta, \\ se_{2n+2}(\eta) &= \sum_{r=0}^{\infty}(-1)^{n+r}B_{2r+2}^{(2n+2)}\sin(2r+2)\eta, \\ se_{2n+1}(\eta) &= \sum_{r=0}^{\infty}(-1)^{n+r}A_{2r+1}^{(2n+1)}\sin(2r+1)\eta. \end{aligned}\right\} \tag{2.16}$$

Of the corresponding solutions of equation (2.14), Davies [2] adopted the functions $Cek_n(\xi)$ and $Sek_n(\xi)$ which are expressed in series of the modified Bessel function $K_r(k\cosh\xi)$ with argument $k\cosh\xi$ and which tend to zero as ξ tends to infinity. Since, however, these series converge rather slowly, it is preferable to make use of the functions $Fek_n(\xi)$ and $Gek_n(\xi)$ which are expressed in series of products of the modified Bessel functions K_r and I_r. These modified Bessel function product series converge more rapidly than the above-mentioned modified Bessel function series with argument $k\cosh\xi$; the rate of convergence increases with increase in ξ (real), and also they converge uniformly in any finite region of the ξ-plane, including the origin.

Thus, we use here, as in our papers [5, 6], the functions $FEK_n(\xi)$ and $GEK_n(\xi)$, which differ from $Fek_n(\xi)$ and $Gek_n(\xi)$ respectively by certain constant multipliers, and which correspond to $ce_n(\eta)$ and $se_n(\eta)$ respectively. These functions are expressed in the modified Bessel function product series as follows:

$$\left.\begin{aligned} FEK_{2n}(\xi) &= \frac{1}{A_0^{(2n)}}\sum_{r=0}^{\infty}A_{2r}^{(2n)}I_r(v_1)K_r(v_2), \\ FEK_{2n+1}(\xi) &= \frac{1}{B_1^{(2n+1)}}\sum_{r=0}^{\infty}B_{2r+1}^{(2n+1)}\{I_r(v_1)K_{r+1}(v_2) - I_{r+1}(v_1)K_r(v_2)\}, \\ GEK_{2n+1}(\xi) &= \frac{1}{A_1^{(2n+1)}}\sum_{r=0}^{\infty}A_{2r+1}^{(2n+1)}\{I_r(v_1)K_{r+1}(v_2) + I_{r+1}(v_1)K_r(v_2)\}, \\ GEK_{2n+2}(\xi) &= \frac{1}{B_2^{(2n+2)}}\sum_{r=0}^{\infty}B_{2r+2}^{(2n+2)}\{I_r(v_1)K_{r+2}(v_2) - I_{r+2}(v_1)K_r(v_2)\}, \end{aligned}\right\} \tag{2.17}$$

with $v_1 = \frac{1}{2}ke^{-\xi}$, $v_2 = \frac{1}{2}ke^{\xi}$.

Thus, we have ultimately appropriate fundamental solutions of the differential equation (2.4) in the forms:

$$\left.\begin{aligned} F_n &= e^{k\cosh\xi\cos\eta} FEK_n(\xi)ce_n(\eta), \qquad (n = 0, 1, 2, \cdots) \\ G_n &= e^{k\cosh\xi\cos\eta} GEK_n(\xi)se_n(\eta). \qquad (n = 1, 2, \cdots) \end{aligned}\right\} \tag{2.18}$$

3. The general expression for the skin friction of the flat plate. Next, we shall proceed to the calculation of the general expression for the skin friction of the flat plate. In general, the drag on a cylinder with arbitrary cross-section is obtained by summing up viscous stresses acting on the surface of the cylinder. Thus, denoting the drag per unit span of the cylinder by D,

$$D = \int_s (p_{xx} l + p_{xy} m)\, ds, \tag{3.1}$$

where ds is the line-element of the circumferential curve s of the cylinder and (l, m) are the direction-cosines of the outward normal to s.

Using

$$p_{xx} = -p + 2\rho\nu \frac{\partial u}{\partial x}, \qquad p_{xy} = \rho\nu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right),$$

the expression for D becomes

$$D = \int_s \left(-p + 2\rho\nu \frac{\partial u}{\partial x}\right) l\, ds + \rho\nu \int_s \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) m\, ds. \tag{3.2}$$

At the surface of the cylinder $u = 0$ and $v = 0$, so that we have, from the second equation in (2.2),

$$p = \rho P. \tag{3.3}$$

The fluid being assumed to be incompressible, the equation of continuity is $\partial u/\partial x + \partial v/\partial y = 0$. Taking account of this equation, together with

$$\zeta = \partial v/\partial x - \partial u/\partial y,$$

the expression for D can be transformed into

$$D = -\rho \int_s Pl\, ds - \rho\nu \int_s \zeta m\, ds - 2\rho\nu \int_s \left(\frac{\partial v}{\partial y} l - \frac{\partial v}{\partial x} m\right) ds. \tag{3.4}$$

But, it is readily seen that

$$\int_s \left(\frac{\partial v}{\partial y} l - \frac{\partial v}{\partial x} m\right) ds = \int_s \left(\frac{\partial v}{\partial x} dx + \frac{\partial v}{\partial y} dy\right) = \int_s dv = 0,$$

and therefore, the expression for D reduces further to

$$D = -\rho \int_s (Pl + \nu\zeta m)\, ds = -\rho \int_s (P\, dy - \nu\zeta\, dx). \tag{3.5}$$

Now, in the case of the flat plate under consideration, α and β are connected

with x and y by the relation (2.3) and consequently, if, in place of (α, β), we use (x, y) as independent variables, the first equation in (2.1) becomes

$$\frac{\partial P}{\partial x} + \nu \frac{\partial \zeta}{\partial y} = 0.$$

Integrating the left-hand side over the region S bounded internally by the cylinder s and externally by an arbitrary contour σ contained in the fluid, we have

$$\iint_S \left(\frac{\partial P}{\partial x} + \nu \frac{\partial \zeta}{\partial y}\right) dx\, dy = 0.$$

An application of Green's theorem to this double integral gives

$$\iint_S \left(\frac{\partial P}{\partial x} + \nu \frac{\partial \zeta}{\partial y}\right) dx\, dy = \int_s (P\, dy - \nu\zeta\, dx) - \int_\sigma (P\, dy - \nu\zeta\, dx),$$

and consequently we have

$$\int_s (P\, dy - \nu\zeta\, dx) = \int_\sigma (P\, dy - \nu\zeta\, dx). \tag{3.6}$$

Therefore, the expression (3.5) for D can be rewritten as:

$$D = -\rho \int_\sigma (P\, dy - \nu\zeta\, dx). \tag{3.7}$$

In our case of the flat plate, we have, by (2.3) and (2.6),

$$P = -\nu \frac{\partial \Phi}{\partial \beta} = \frac{\nu}{U} \frac{\partial \Phi}{\partial y}, \qquad \zeta = \frac{\partial \Phi}{\partial \alpha} = -\frac{1}{U} \frac{\partial \Phi}{\partial x}.$$

Thus, the general expression for the skin friction D of the flat plate becomes ultimately

$$D = -\frac{\rho\nu}{U} \int_\sigma \left(\frac{\partial \Phi}{\partial x}\, dx + \frac{\partial \Phi}{\partial y}\, dy\right) = -\frac{\rho\nu}{U} \int_\sigma d\Phi = -\frac{\rho\nu}{U} [\Phi]_\sigma, \tag{3.8}$$

where $[\Phi]_\sigma$ denotes the change in Φ when taken once round the closed contour σ. It is seen that the only contribution to the skin friction D of the flat plate arises from many-valued solutions of equation (2.4). This result is due originally to Davies [2].

4. The stream function. The stream function ψ for the flow past the flat plate under consideration is considered to be composed of the stream function β for the potential flow and an additional part, ψ_1 say, due to the effect of the viscosity of the fluid. Thus,

$$\psi = \beta + \psi_1. \tag{4.1}$$

The additional stream function ψ_1 is connected with the vorticity ζ as:

$$\frac{\partial^2 \psi_1}{\partial x^2} + \frac{\partial^2 \psi_1}{\partial y^2} = \zeta. \tag{4.2}$$

Hence, taking (2.6), (2.4), and (2.3) into account, we have

$$\frac{\partial^2\psi_1}{\partial x^2}+\frac{\partial^2\psi_1}{\partial y^2}=\frac{\partial\Phi}{\partial\alpha}=-\nu\left(\frac{\partial^2\Phi}{\partial\alpha^2}+\frac{\partial^2\Phi}{\partial\beta^2}\right)=-\frac{\nu}{U^2}\left(\frac{\partial^2\Phi}{\partial x^2}+\frac{\partial^2\Phi}{\partial y^2}\right),$$

or

$$\nabla^2\psi_1=-(\nu/U^2)\nabla^2\Phi,$$

where ∇^2 stands for $\partial^2/\partial x^2+\partial^2/\partial y^2$.

This gives

$$\psi_1=(\text{a harmonic function})-(\nu/U^2)\Phi, \tag{4.3}$$

and therefore, since $\beta=-Uc\sinh\xi\sin\eta$ by (2.11), the expression for ψ becomes

$$\psi=-Uc\sinh\xi\sin\eta+(\text{a harmonic function})-(\nu/U^2)\Phi. \tag{4.4}$$

Evidently ψ must be an odd function with respect to η, so that we may write

$$(\text{a harmonic function})=A_0\eta+\textstyle\sum_{n=1}^{\infty}A_n e^{-n\xi}\sin n\eta, \tag{4.5}$$

where $A_0, A_1, A_2, \cdots$ are constants.

As mentioned at the end of the last section, the function Φ must be a many-valued function of η. On the other hand, the stream function ψ itself must be a periodic function of η with period 2π. Therefore, when the expression (4.5) is substituted into (4.4), a many-valued term in Φ must cancel the many-valued term $A_0\eta$ there. Hence it will readily be seen that a linear combination of the fundamental odd solutions G_n's ($n=1, 2, 3, \cdots$) as given in (2.18) cannot be adopted for the function Φ.

We now introduce functions Φ_m's such that

$$\Phi_m=\int_0^{\eta}\left\{\frac{\partial F_m}{\partial\xi}-2kF_m\sinh\xi\cos\eta\right\}d\eta, \qquad (m=0, 1, 2, \cdots) \tag{4.6}$$

where F_m's ($m=0, 1, 2, \cdots$) are the fundamental even solutions as given in (2.18). Evidently such functions Φ_m's are all odd functions of η and also it is readily shown as follows that they satisfy equation (2.4), which becomes, when the elliptic coordinates (ξ, η) are used as independent variables,

$$\frac{\partial^2\Phi}{\partial\xi^2}+\frac{\partial^2\Phi}{\partial\eta^2}=2k\left(\frac{\partial\Phi}{\partial\xi}\sinh\xi\cos\eta-\frac{\partial\Phi}{\partial\eta}\cosh\xi\sin\eta\right). \tag{4.7}$$

From (4.6) we have

$$\frac{\partial\Phi_m}{\partial\xi}=\int_0^{\eta}\left\{\frac{\partial^2F_m}{\partial\xi^2}-2k\frac{\partial F_m}{\partial\xi}\sinh\xi\cos\eta-2kF_m\cosh\xi\cos\eta\right\}d\eta. \tag{4.8}$$

Making use of the fact that F_m satisfies equation (2.4), and hence also (4.7), we find, on substituting for $\partial^2F_m/\partial\xi^2$, that the right-hand side of (4.8) is immediately integrable, giving

$$\frac{\partial\Phi_m}{\partial\xi}=-\left(\frac{\partial F_m}{\partial\eta}+2kF_m\cosh\xi\sin\eta\right). \tag{4.9}$$

On the other hand, it follows immediately from (4.6) that

$$\frac{\partial \Phi_m}{\partial \eta} = \frac{\partial F_m}{\partial \xi} - 2kF_m \sinh \xi \cos \eta. \tag{4.10}$$

Hence we readily find that

$$\frac{\partial^2 \Phi_m}{\partial \xi^2} + \frac{\partial^2 \Phi_m}{\partial \eta^2} = 2k\left(\frac{\partial \Phi_m}{\partial \xi} \sinh \xi \cos \eta - \frac{\partial \Phi_m}{\partial \eta} \cosh \xi \sin \eta\right),$$

namely, the functions Φ_m's satisfy equation (4.7), and hence also equation (2.4).

Further, it is seen immediately that each of the functions Φ_m's contains a linear term in η.

Thus, we can adequately put

$$-\frac{\nu}{U^2}\Phi = \sum_{m=0}^{\infty} B_m \Phi_m, \tag{4.11}$$

where B_m's are constants, and then we obtain ultimately the expression for the stream function ψ in the form:

$$\psi = -Uc \sinh \xi \sin \eta + A_0\eta + \sum_{n=1}^{\infty} A_n e^{-n\xi} \sin n\eta + \sum_{m=0}^{\infty} B_m \Phi_m. \tag{4.12}$$

5. Boundary conditions. The constants A_0, A_n $(n = 1, 2, 3, \cdots)$, $B_m (m = 0, 1, 2, \cdots)$ contained in (4.12) must be determined by the boundary conditions. Since, however, it is readily seen that the boundary conditions at infinity, i.e. $u = U$, $v = 0$ as $\xi \to \infty$ are already satisfied by the velocity components u, v as obtained from ψ in (4.12), we have only to consider the boundary conditions at the surface of the flat plate ($\xi = 0$), which are expressed as:

$$\frac{\partial \psi}{\partial \xi} = 0, \qquad \frac{\partial \psi}{\partial \eta} = 0 \qquad \text{at } \xi = 0. \tag{5.1}$$

On substituting the expression (4.12) for ψ, these conditions yield

$$\left.\begin{aligned} Uc \sin \eta + \sum_{n=1}^{\infty} nA_n \sin n\eta - \sum_{m=0}^{\infty} B_m \left(\frac{\partial \Phi_m}{\partial \xi}\right)_{\xi=0} = 0, \\ A_0 + \sum_{n=1}^{\infty} nA_n \cos n\eta + \sum_{m=0}^{\infty} B_m \left(\frac{\partial \Phi_m}{\partial \eta}\right)_{\xi=0} = 0. \end{aligned}\right\} \tag{5.2}$$

Making use of the functions F_n's in (2.18), we find, from (4.9) and (4.10), that

$$\left.\begin{aligned} \left(\frac{\partial \Phi_m}{\partial \xi}\right)_{\xi=0} &= -e^{k\cos\eta} FEK_m(0)[ce_m'(\eta) + k\, ce_m(\eta) \sin \eta], \\ \left(\frac{\partial \Phi_m}{\partial \eta}\right)_{\xi=0} &= e^{k\cos\eta} FEK_m'(0) ce_m(\eta), \end{aligned}\right\} \tag{5.3}$$

where dashes denote differentiation with respect to ξ or η.

We expand the right-hand members in (5.3) into Fourier sine and cosine series respectively as follows:

$$\left(\frac{\partial \Phi_m}{\partial \xi}\right)_{\xi=0} = -\sum_{n=1}^{\infty} H_{m,n} \sin n\eta, \qquad \left(\frac{\partial \Phi_m}{\partial \eta}\right)_{\xi=0} = \sum_{n=0}^{\infty} K_{m,n} \cos n\eta. \tag{5.4}$$

Substituting these into the last terms on the left-hand sides of equations (5.2) and equating the coefficients of $\sin n\eta$ and $\cos n\eta$ to zero, we obtain, from the first equation,

$$\left.\begin{aligned} Uc + A_1 + \sum_{m=0}^{\infty} H_{m,1}B_m &= 0, \\ nA_n + \sum_{m=0}^{\infty} H_{m,n}B_m &= 0, \qquad (n = 2, 3, \cdots) \end{aligned}\right\} \tag{5.5}$$

while the second equation yields

$$\left.\begin{aligned} A_0 + \sum_{m=0}^{\infty} K_{m,0}B_m &= 0, \qquad A_1 + \sum_{m=0}^{\infty} K_{m,1}B_m = 0, \\ nA_n + \sum_{m=0}^{\infty} K_{m,n}B_m &= 0. \qquad (n = 2, 3, \cdots) \end{aligned}\right\} \tag{5.6}$$

From the first equation in (5.6) we have

$$A_0 = -\sum_{m=0}^{\infty} K_{m,0}B_m, \tag{5.7}$$

and eliminating A_1 between the first equation in (5.5) and the second equation in (5.6) we obtain

$$\sum_{m=0}^{\infty} (K_{m,1} - H_{m,1})B_m = Uc. \tag{5.8}$$

Also, elimination of A_n between the second equation in (5.5) and the third equation in (5.6) yields

$$\sum_{m=0}^{\infty} (K_{m,n} - H_{m,n})B_m = 0. \qquad (n = 2, 3, \cdots) \tag{5.9}$$

Hence, writing, for brevity,

$$\lambda_{m,n} = K_{m,n} - H_{m,n}, \qquad \begin{pmatrix} m = 0, 1, 2, \cdots \\ n = 1, 2, 3, \cdots \end{pmatrix} \tag{5.10}$$

we have

$$\sum_{m=0}^{\infty} \lambda_{m,n}B_m = \begin{cases} Uc, & (n = 1) \\ 0. & (n = 2, 3, \cdots) \end{cases} \tag{5.11}$$

The constants B_m's ($m = 0, 1, 2, \cdots$) can be determined by solving this infinite system of simultaneous linear equations, and if the values of B_m's so determined are substituted into (5.7), the value of A_0 can be found. Also, the value of A_1 can be obtained either from the first equation in (5.5) or from the second equation in (5.6), and further all the remaining constants A_n's ($n = 2, 3, \cdots$) can be determined either from the second equation in (5.5) or from the third equation in (5.6).

6. Reduction of the expressions for $\lambda_{m,n}$'s. Next we shall reduce the expressions (5.10) for $\lambda_{m,n}$'s to such forms that are more suitable for calculation. For this purpose, we first introduce the quantities $p_{m,n}$'s defined as:

$$p_{m,n} = \frac{1}{\pi}\int_{-\pi}^{\pi} e^{k\cos\eta} ce_m(\eta) \cos n\eta \, d\eta, \qquad (m, n = 0, 1, 2, \cdots). \tag{6.1}$$

Making use of the first and the second equations in (2.16) and carrying out

integrations, we get*

$$\left.\begin{aligned} p_{2m,n} &= \sum_{r=0}^{\infty}(-1)^{m+r}A_{2r}^{(2m)}\{I_{n-2r}(k) + I_{n+2r}(k)\}, \\ p_{2m+1,n} &= \sum_{r=0}^{\infty}(-1)^{m+r}B_{2r+1}^{(2m+1)}\{I_{n-2r-1}(k) + I_{n+2r+1}(k)\}. \end{aligned}\right\} \tag{6.2}$$

If we use these $p_{m,n}$'s, we find, by the first equations in (5.3) and (5.4), that the expressions for $H_{m,n}$'s become

$$\begin{gathered} H_{m,n} = FEK_m(0)\{kp_{m,n-1} - np_{m,n} - kp_{m,n+1}\}, \\ (m = 0, 1, 2, \cdots; n = 1, 2, 3, \cdots) \end{gathered} \tag{6.3}$$

or, if, for convenience's sake, we put

$$q_{m,n} = kp_{m,n-1} - np_{m,n} - kp_{m,n+1}, \tag{6.4}$$

we have

$$H_{m,n} = FEK_m(0)q_{m,n}. \qquad \begin{pmatrix} m = 0, 1, 2, \cdots \\ n = 1, 2, 3, \cdots \end{pmatrix} \tag{6.5}$$

In like manner we obtain from the second equations in (5.3) and (5.4) the expressions for $K_{m,n}$'s as follows:

$$\left.\begin{aligned} K_{m,0} &= \tfrac{1}{2}FEK_m'(0)p_{m,0}, \qquad (m = 0, 1, 2, \cdots) \\ K_{m,n} &= FEK_m'(0)p_{m,n}. \qquad \begin{pmatrix} m = 0, 1, 2, \cdots \\ n = 1, 2, 3, \cdots \end{pmatrix} \end{aligned}\right\} \tag{6.6}$$

Thus, we find that the expressions for $\lambda_{m,n}$'s ($n \geqq 1$) are ultimately reduced to

$$\begin{gathered} \lambda_{m,n} = FEK_m'(0)p_{m,n} - FEK_m(0)q_{m,n}. \\ (m = 0, 1, 2, \cdots; n = 1, 2, 3, \cdots) \end{gathered} \tag{6.7}$$

7. Expansion formula for the skin friction of the flat plate. The general expression for the skin friction D of the flat plate is given by (3.8), namely:

$$D = -\frac{\rho\nu}{U}[\Phi]_\sigma, \tag{7.1}$$

where $[\Phi]_\sigma$ is the change in Φ when taken once round the closed contour σ.

Since, as mentioned already, the many-valued term in $\sum_{m=0}^{\infty}B_m\Phi_m$

$$(= -(\nu/U^2)\Phi)$$

in the expression (4.12) for ψ must be equal to $-A_0\eta$ there, we easily find that

$$[\Phi]_\sigma = 2\pi A_0\frac{U^2}{\nu}. \tag{7.2}$$

* Here use has been made of the fact that when s is an integer, the modified Bessel function $I_s(x)$ is expressed as:

$$I_s(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{x\cos\theta}\cos s\theta\, d\theta.$$

Cf. G. N. Watson, '*Theory of Bessel functions*' (Cambridge Univ. Press, 2nd ed. 1948, p. 181).

Hence we have

$$D = -2\pi\rho U A_0 , \tag{7.3}$$

which, by means of (5.7), becomes

$$D = 2\pi\rho U \sum_{m=0}^{\infty} K_{m,0} B_m . \tag{7.4}$$

Bearing in mind that the length of the flat plate is $2c$, we define, as usual, the skin friction coefficient C_D as:

$$C_D = D/(\rho U^2 \cdot 2c).$$

Then we have

$$C_D = \frac{2\pi}{\nu R} \sum_{m=0}^{\infty} K_{m,0} B_m , \tag{7.5}$$

where R is the Reynolds number, namely $R = U \cdot 2c/\nu = 4k$. This is the general expression for the skin friction coefficient of the flat plate.

Next, we shall proceed to the calculation of an expansion formula for C_D, assuming the Reynolds number R $(=4k)$ to be sufficiently small. However, we here confine ourselves to the deduction of an expansion formula correct to the order of R^2.

To do this, it is sufficient to retain only the first two terms in the series

$$\sum_{m=0}^{\infty} K_{m,0} B_m$$

in (7.5), and for determining the values of B_0 and B_1, it is sufficient to solve the following system of simultaneous linear equations:

$$\left.\begin{aligned} \lambda_{0,1} B_0 + \lambda_{1,1} B_1 + \lambda_{2,1} B_2 &= Uc, \\ \lambda_{0,2} B_0 + \lambda_{1,2} B_1 + \lambda_{2,2} B_2 &= 0, \\ \lambda_{0,3} B_0 + \lambda_{1,3} B_1 + \lambda_{2,3} B_2 &= 0, \end{aligned}\right\} \tag{7.6}$$

which have been obtained from the first three equations in (5.11) by putting $B_3 = B_4 = \cdots = 0$.

Expansions in power series of the Reynolds number R for the nine $\lambda_{m,n}$'s $(m = 0, 1, 2;\ n = 1, 2, 3)$ necessary for our calculation have been obtained by expanding various quantities such as $FEK_m(0)$, $FEK_m'(0)$, $p_{m,n}$, and $q_{m,n}$ in power series of R. For shortness the results only will be given here.

Thus, writing $q = \frac{1}{8}R$, we have

$$\lambda_{0,1} = -2(1 + S_1)q - \tfrac{1}{2}(1 + S_1)q^3 + O(q^5),$$

$$\lambda_{0,2} = -\tfrac{1}{2}(3 + 2S_1)q^2 + O(q^4),$$

$$\lambda_{0,3} = -\tfrac{1}{6}(5 + 3S_1)q^3 + O(q^5);$$

$$\lambda_{1,1} = -\tfrac{1}{4}(7 + 4S_1)q + O(q^3),$$

$$\lambda_{1,2} = -1 + O(q^2), \qquad \lambda_{1,3} = -\tfrac{3}{4}q + O(q^3);$$

$$\lambda_{2,1} = \frac{2}{q} + O(q), \qquad \lambda_{2,2} = -\tfrac{4}{3} + O(q^2), \qquad \lambda_{2,3} = -\frac{2}{q} + O(q),$$

where

$$S_1 = -\gamma - \log \tfrac{1}{2}q = -\gamma - \log \tfrac{1}{16}R,$$

$\gamma = 0.57721\ldots$ being Euler's constant.

Solving equations (7.6) and making use of the above expansions, we have

$$\left.\begin{aligned} B_0 &= -\frac{2\nu}{1+S_1}\left\{1 + \frac{1}{24(1+S_1)}(29 + 36S_1 + 12S_1^2)q^2\right\} + O(q^4), \\ B_1 &= \frac{\nu}{1+S_1}(3 + 2S_1)q^2 + O(q^4). \end{aligned}\right\} \tag{7.7}$$

Also, from the first equation in (6.6) we obtain the expansions for $K_{0,0}$ and $K_{1,0}$ as:

$$K_{0,0} = -\{1 + \tfrac{1}{2}q^2 + O(q^4)\}, \qquad K_{1,0} = -\{1 + \tfrac{1}{8}q^2 + O(q^4)\}. \tag{7.8}$$

Thus, we have

$$\begin{aligned} \sum_{m=0}^{\infty} K_{m,0}B_m &= K_{0,0}B_0 + K_{1,0}B_1 + \cdots \\ &= \frac{2\nu}{1+S_1}\left\{1 - \frac{1}{2(1+S_1)}(S_1^2 + S_1 - \tfrac{5}{12})q^2\right\} + O(q^4), \end{aligned} \tag{7.9}$$

and therefore, by (7.5), a second approximation correct to the order of R^2 for the skin friction coefficient C_D of the flat plate becomes ultimately

$$\left.\begin{aligned} C_D &= \frac{4\pi}{R(1+S_1)}\left[1 - \frac{1}{1+S_1}(S_1^2 + S_1 - \tfrac{5}{12})\frac{R^2}{128}\right], \\ &\text{where} \\ S_1 &= -\gamma - \log \tfrac{1}{16}R. \end{aligned}\right\} \tag{7.10}$$

This is just what has been obtained by Piercy and Winny [1], pursuing Bairstow, Cave, and Lang's analysis [8].

If, in place of S_1, we use S such that $S = 1 + S_1$, the above formula may also be written in the form:

$$\left.\begin{aligned} C_D &= \frac{4\pi}{RS}\left[1 - \frac{1}{S}(S^2 - S - \tfrac{5}{12})\frac{R^2}{128}\right], \\ &\text{with} \\ S &= 1 - \gamma - \log \tfrac{1}{16}R. \end{aligned}\right\} \tag{7.11}$$

Thus it has been ascertained that contrary to Davies's argument, an expansion formula in complete agreement with Piercy and Winny's can still be derived for the skin friction coefficient of the flat plate by pursuing Davies's original work, if, in place of $Cek_n(\xi)$ and $Sek_n(\xi)$ used by himself, we make use of $FEK_n(\xi)$ and $GEK_n(\xi)$, which differ from $Fek_n(\xi)$ and $Gek_n(\xi)$ respectively by certain constant multipliers, for solutions of the modified Mathieu equation, these functions being expressed in more rapidly convergent series in products of the modified Bessel functions.

REFERENCES

1. N. A. V. PIERCY AND H. F. WINNY, The skin friction of flat plates to Oseen's approximation. Proc. Roy. Soc. A, **140** (1933), 543.
2. T. V. DAVIES, An investigation of the flow of a viscous fluid past a flat plate, using elliptic coordinates. Phil. Mag. **31** (1941), 283.
3. S. SIDRAK, The flow of a viscous liquid past an elliptic cylinder. Proc. Roy. Irish Acad. **53** (1950), 65.
4. S. TOMOTIKA, A note on the skin friction of a flat plate to Oseen's approximation. Mem. Coll. Sci. Univ. of Kyoto, A, **27** (1953), 1.
5. S. TOMOTIKA AND T. AOI, The steady flow of a viscous fluid past an elliptic cylinder and a flat plate at small Reynolds numbers. Quart. J. Mech. and Applied Math. **6** (1953), 290.
6. S. TOMOTIKA AND H. YOSINOBU, The drag on an elliptic cylinder moving in a viscous liquid at small Reynolds numbers. Mem. Coll. Sci. Univ. of Kyoto, A, **27** (1953), 11.
7. T. LEWIS, Solutions of Oseen's extended equations for circular and elliptic cylinders and a flat plate. Quart. J. Math., Oxford Series, **9** (1938), 21.
8. L. BAIRSTOW, B. M. CAVE, AND E. D. LANG, The resistance of a cylinder moving in a viscous fluid. Phil. Trans. Roy. Soc. A, **223** (1923), 383.

UNIVERSITY OF KYOTO,
KYOTO, JAPAN.

(Received February 8, 1955)

SMALL PERTURBATIONS IN THE UNSTEADY FLOW OF A COMPRESSIBLE, VISCOUS AND HEAT-CONDUCTING FLUID*

By T. Yao-tsu Wu

I. Introduction. In the present work interest is focused on the flow of a compressible, viscous and heat-conducting fluid resulting from certain small external forces and heat additions. It is expected that the responses of the flow field are also small perturbations from the basic flow. Consequently, as a first order approximation, the equations governing the fluid motion can be linearized. Then the analysis is simplified enough to allow a unified discussion of the various effects. Limiting solutions of this linearized system of equations exhibit automatically the regions where viscosity and heat conduction play significant roles. In contrast to the simplification of the boundary layer type, the present theory gives more of the over-all picture while some of the typical features of the more exact nonlinear problems are still retained. Furthermore, within the frame of linear theory, superposition can always be applied to construct new solutions. These viewpoints are also explained in some previous works[1-4], where the flow of real fluid is investigated for various idealizations. In reference [1]† a detailed study is made of the propagation of small disturbances in a compressible, viscous but non-heat-conducting fluid. A similar analysis is carried out in Ref. [4] for the problem of turbulence in a compressible, viscous and heat-conducting fluid.

With respect to this linearized system of equations the concepts of longitudinal and transversal waves are generalized to include heat conduction‡. A proof is given for the statement that any linearized flow may be split uniquely into a longitudinal and a transversal component.

It follows from the generalized definition of these waves that the compressibility and heat-conducting effects appear only in the longitudinal part. The transversal wave, which has been thoroughly discussed in previous works[1,4], is not considered any further in this paper. As a simple example of the generalized longitudinal wave, the one-dimensional nonstationary flow is first studied. The equations, after linearization, are further simplified with the assumption that the Prandtl number of the fluid is equal to $\frac{3}{4}$. The corresponding fundamental solutions are found and are approximated for both large and small values of time. Some interesting features shown by these solutions are discussed in detail.

The author wishes to thank Professors J. D. Cole and P. A. Lagerstrom for their valuable suggestions.

* This work was carried out under the Office of Naval Research Contract N6 ONR-244, Task Order VIII, Project NR 061-036.

† An extensive bibliography of related topics has been collected in Ref. [1].

‡ Reference [1] is referred to for the concepts of these waves when heat conduction is omitted.

II. Linearization of Basic Equations; Theorem of Splitting. The medium is assumed to be a perfect gas which fills a space of infinite extent and obeys the equation of state

$$P = R\rho T \tag{1}$$

where P is the pressure, R the gas constant, ρ the density and T the temperature. Now our purpose is to investigate the resulting flow field P, ρ, T, and the flow velocity $\vec{q}$, as functions of space and time, due to a distribution of external force $\vec{F}$ (per unit mass) and heat addition Q (per unit mass per unit time); the initial conditions being uniform and quiescent: $P = P_0$, $\rho = \rho_0$, $T = T_0$ and $\vec{q} = 0$. The basic equations governing the motion of such a compressible, viscous and heat-conducting fluid can be derived by applying the principles of conservation of mass, momentum and energy. They can be found elsewhere in the books on fluid dynamics (e.g. Ref. 5, p. 49) and hence will not be reproduced here. These three equations together with the equation of state, Eq. (1), are a system of non-linear differential equations for four unknowns, $\vec{q}$, P, ρ and T, provided that the coefficient of viscosity μ, the heat conductivity k and the specific heats are known functions of T.

While the general solution to this complicated system is still unattainable, the mathematical characteristics of this nonlinear system, however, usually indicate some basic features of the system by showing regions in which certain disturbances (or discontinuities) may propagate. These features should remain even when approximations, such as linearization, are introduced. When both μ and k are different from zero, it can be shown[1] that time t equal to constant is a characteristic. This set indicates propagation of disturbances with an infinite speed and thus represents the usual diffusive behavior of viscous and heat-conducting effects by which any discontinuity will instantly be smoothed out.

However, the characteristic condition changes when both viscosity and heat conduction are neglected. In the case of one-dimensional flow, $\vec{q} = (u, 0, 0)$ where u and the thermodynamic variables depend only on the space coordinate x and time t. Then when $\mu = 0$, $k = 0$, the mathematical characteristics $\psi(x, t) =$ constant are given by[1]

$$(\psi_t + u\psi_x)[\psi_t + (u \pm c)\psi_x] = 0 \tag{2}$$

where $c = (\gamma RT)^{\frac{1}{2}}$, the usual adiabatic speed of sound. The first bracket indicates the streamlines of the flow across which temperature, density and velocity may jump (such as slip stream and interface). The second bracket indicates propagation of disturbances with sound speed c relative to the fluid. These characteristic lines for μ, $k = 0$ must play a significant role in the solutions for small values of μ and k. This remark will be clarified later in the discussion of solutions of our linearized system (see the last section and the Appendix).

In the following the analysis will be based on the theory of small perturbations. By small disturbances we mean that the applied force $\vec{F}$ and heat addition Q are small and they vanish sufficiently fast at infinity. Then we may assume

small perturbations in the resulting flow:

$$P = P_0(1 + p), \qquad \rho = \rho_0(1 + s), \qquad T = T_0(1 + \theta), \tag{3}$$

where almost everywhere

$$p, \theta, s \ll 1 \quad \text{and} \quad |\vec{q}| \ll c. \tag{3a}$$

It is also assumed that

$$\mu = \mu_0(1 + \mu'), \qquad k = k_0(1 + k'), \qquad \mu', k' \ll 1 \tag{4}$$

where μ_0 and k_0 depend on T_0. Substituting Eqs. (3), (4) into the original nonlinear equations and neglecting all squared terms of small quantities, we obtain a linearized system of equations:

$$p = s + \theta, \tag{5}$$

$$s_t + \operatorname{div} \vec{q} = 0, \tag{6}$$

$$\vec{q}_t + (c^2/\gamma) \operatorname{grad} p = \vec{F} + \tfrac{4}{3}\nu \operatorname{grad} \operatorname{div} \vec{q} - \operatorname{curl} \operatorname{curl} \vec{q}, \tag{7}$$

$$\theta_t - \kappa\Delta\theta = (\gamma - 1)s_t + \Omega \tag{8}$$

where Δ stands for Laplacian, $\gamma = c_p/c_v$ (ratio of the specific heats), $\nu = \mu_0/\rho_0$ (kinematic viscosity of the undisturbed flow), $\kappa = k_0/(c_v\rho_0)$ (thermometric conductivity) and $\Omega = Q/(c_vT_0)$, the non-dimensional heat addition per unit time. In the momentum equation (7) the Stokes assumption on the second coefficient of viscosity is adopted. In the energy equation (8) the dissipation function is dropped out by linearization because it contains only squared terms of small quantities. Thus the viscous effects do not influence the linearized energy relation directly. Furthermore, the effects of $\vec{F}$ and Ω become independent of each other when the dissipation is neglected. Consequently, the linearization procedure will not yield a good approximation for those problems in which the dissipation is an important factor in the energy equation.

In the following it will be shown that a flow governed by the linear system (5)–(8) can be represented uniquely as the sum of a longitudinal wave (having an irrotational velocity field) and a transversal wave (having a solenoidal velocity field). Such a splitting of the flow field is of fundamental importance. It gives some clue as to which part of the more exact nonlinear flow the vorticity, pressure and temperature variations should be significantly associated. Splitting of incompressible waves has been discussed by Lamb[6] and Oseen[7]. For compressible, viscous but non-heat-conducting fluid the splitting has been studied by Lagerstrom, Cole and Trilling[1]. A similar splitting for compressible, viscous and heat-conducting fluid has been introduced by Kovasznay[4].

Now we shall generalize the original definition[1] of longitudinal and transversal waves to include heat conduction.

1. *Transversal Wave.* A necessary condition for a wave to be transversal is that $\operatorname{div} \vec{q} = 0$. The continuity equation (6) then becomes $s_t = 0$. Hence we

may assert that $s = 0$ if s vanishes initially. We further postulate that $p = \theta = 0$. It then follows from Eqs. (7) and (8) that div $\vec{F} = 0$ and $\Omega = 0$. Thus the transversal waves are defined as:

$$\vec{q}_t = -\nu \operatorname{curl} \operatorname{curl} \vec{q} + \vec{F} \equiv \nu \Delta \vec{q} + \vec{F}, \tag{9a}$$

$$\operatorname{div} \vec{q} = 0, \qquad \operatorname{div} \vec{F} = 0, \qquad p = s = \theta = 0, \qquad \Omega = 0. \tag{9b}$$

Apparently, the transversal waves, generated by solenoidal force field, do not contain compressibility effect and heat conduction, and have no first order interaction with pressure and temperature fields. Taking the curl of Eq. (9a), one obtains the familiar equation that governs the diffusion of vorticity.

2. *Longitudinal Wave.* For a longitudinal wave, curl $\vec{q} = 0$. It then follows from Eq. (7) that a consistent force field must also be irrotational, that is, curl $\vec{F} = 0$. Except for these two conditions there are no other restrictions on the variables. Hence the longitudinal waves can be defined by:

$$\vec{q}_t + (c^2/\gamma)\nabla p = \tfrac{4}{3}\nu \operatorname{grad} \operatorname{div} \vec{q} + \vec{F} \equiv \tfrac{4}{3}\nu \Delta \vec{q} + \vec{F}, \tag{10a}$$

$$s_t + \operatorname{div} \vec{q} = 0, \tag{10b}$$

$$\theta_t - \kappa \Delta \theta = (\gamma - 1)s_t + \Omega, \tag{10c}$$

$$p = s + \theta, \tag{10d}$$

$$\operatorname{curl} \vec{q} = 0, \qquad \operatorname{curl} \vec{F} = 0. \tag{10e}$$

Furthermore, Eq. (10e) implies that there always exist a scalar velocity potential ϕ and a force potential Ξ such that

$$\vec{q} = \operatorname{grad} \phi, \qquad \vec{F} = \operatorname{grad} \Xi. \tag{10f}$$

Thus, in a linearized fluid flow, heat conduction and variations in thermodynamic variables are carried by the longitudinal part only while the vorticity is associated solely with the transversal part. It should be remarked that the longitudinal and transversal parts individually represent flows.

Taking the divergence of (10a) and eliminating $\vec{q}$ and θ by using Eqs. (10b, c, d), one obtains

$$p_{tt} - c^2 \Delta p - \kappa \Delta p_t = \kappa(\tfrac{4}{3}Pr - 1)\Delta s_t + \Omega_t - \gamma \Delta \Xi \tag{11}$$

where Pr is the Prandtl number,

$$Pr = c_p \mu_0 / k_0 = \gamma \nu / \kappa.$$

If s is further eliminated the order of the differential equation for p is raised to five. However, it is of interest to note that when the Prandtl number $Pr = \frac{3}{4}$, the term containing s in Eq. (11) vanishes; or

$$p_{tt} - c^2 \Delta p - \kappa \Delta p_t = \Omega_t - \gamma \Delta \Xi. \tag{12}$$

Also, the equations for other dependent variables can be further simplified when $Pr = \frac{3}{4}$ (see the next section). This value of the Prandtl number is actually very

close to the observed value (for air, $Pr = 0.72$ to 0.76 over a wide range of temperature). When Pr is different from $\frac{3}{4}$, one can find that in the equation for p, the coefficients of the highest derivatives $\Delta\Delta p_t$ and Δp_{tt} are both proportional to Pr. Then it follows from the theory of differential equations (Ref. 8, p. 142) that the solutions and their derivatives will be continuous in Pr for $0 < Pr < \infty$; and the solutions for Pr slightly different from $\frac{3}{4}$ can be obtained by a perturbation procedure in Pr. Thus, we see that for $Pr = \frac{3}{4}$ the pressure p satisfies a parabolic equation of third order. It is basically diffusive in nature and therefore its solutions do not admit any discontinuity. However, the value of κ/c^2 is quite small for all practical cases. (For atmospheric air it is of the order of 10^{-9} sec.) Hence in regions where the term $(\kappa/c^2)\Delta p_t$ is small Eq. (12) behaves more like a wave equation; this term then represents the sound absorption and attenuation which become important only at high frequencies.

With the aid of these concepts we may now formulate the "theorem of splitting" for compressible, viscous and heat-conducting fluid flows in the frame of linearized theory as follows:

3. *Theorem of Splitting.* A flow $(\vec{q}, p, s, \theta)$ governed by the linear system (5)–(8) with given Ω and $\vec{F} = \vec{F}_1 + \vec{F}_2$, where curl $\vec{F}_1 = 0$ and div $\vec{F}_2 = 0$, may be expressed uniquely as the sum of a longitudinal wave $(\vec{q}_1, p, s, \theta)$ and a transversal wave $(\vec{q}_2, 0, 0, 0)$ such that $(\vec{q}_1, p, s, \theta)$ with $\vec{F}_1$ and Ω satisfy (10), $(\vec{q}_2, 0, 0, 0)$ with $\vec{F}_2$ and $\Omega = 0$ satisfy (9) and $\vec{q} = \vec{q}_1 + \vec{q}_2$.

Proof: It is known that a vector field $\vec{q}$ may be expressed in more than one way as the sum of an irrotational component $\vec{w}_1$ and a solenoidal component $\vec{w}_2$. That is

$$\vec{q}(\vec{x}, t) = \vec{w}_1(\vec{x}, t) + \vec{w}_2(\vec{x}, t) \tag{13a}$$

where

$$\text{curl } \vec{w}_1 = 0, \qquad \text{div } \vec{w}_2 = 0. \tag{13b}$$

Then Eq. (7) can be written as

$$\partial\vec{w}_1/\partial t + (c^2/\gamma)\nabla p - (4\nu/3)\Delta\vec{w}_1 - \vec{F}_1 = -(\partial\vec{w}_2/\partial t - \nu\Delta\vec{w}_2 - \vec{F}_2). \tag{14}$$

The splitting given by Eq. (13) is unique only up to a harmonic field. In other words, if $\vec{h}$ is a harmonic function so that curl $\vec{h}$ = div $\vec{h}$ = 0, then for the decomposition given by Eq. (13),

$$\vec{q} = \vec{q}_1 + \vec{q}_2 \tag{15a}$$

where

$$\vec{q}_1 = \vec{w}_1 + \vec{h}, \qquad \vec{q}_2 = \vec{w}_2 - \vec{h} \tag{15b}$$

gives another possible decomposition. However, for an initial choice of $\vec{w}_1$ and $\vec{w}_2$ one can always make $(\vec{q}_1, p)$ satisfy Eq. (10a) and make $\vec{q}_2$ satisfy Eq. (9a) by assigning to $\vec{h}$ the value

$$\vec{h} = -\int_{-\infty}^{t} \{\partial\vec{w}_1/\partial\tau + (c^2/\gamma)\nabla p - (4\nu/3)\Delta\vec{w}_1 - \vec{F}_1\}\, d\tau \tag{16}$$

which may also be expressed, by using Eq. (14), as

$$\vec{h} = \int_{-\infty}^{t} \{\partial \vec{w}_2/\partial \tau - \nu \Delta \vec{w}_2 - \vec{F}_2\}\, d\tau \tag{17}$$

That $\vec{h}$ defined above is harmonic follows from Eq. (13b) by taking the curl of Eq. (16) and the divergence of Eq. (17). Now, that

$$\frac{\partial \vec{q}_1}{\partial t} + \frac{c^2}{\gamma} \nabla p - \frac{4\nu}{3} \Delta \vec{q}_1 - \vec{F}_1 = \frac{\partial \vec{w}_1}{\partial t} + \frac{c^2}{\gamma} \nabla p - \frac{4\nu}{3} \Delta \vec{w}_1 - \vec{F}_1 + \frac{\partial \vec{h}}{\partial t} = 0 \tag{18}$$

follows from Eqs. (15b), (16); and from Eqs. (15b) and (17),

$$\frac{\partial \vec{q}_2}{\partial t} - \nu \Delta \vec{q}_2 - \vec{F}_2 = 0. \tag{19}$$

Besides, Eq. (6) may also be written as

$$s_t + \operatorname{div} \vec{q}_1 = 0. \tag{20}$$

Combining Eqs. (18), (20), (5) and (8), one can easily see that $(\vec{q}_1, p, s, \theta)$ with $\vec{F} = \vec{F}_1$ satisfy the longitudinal wave equation (10). Equation (19) also shows that $(\vec{q}_2, 0, 0, 0)$ satisfies the transversal wave equation (9) with $\vec{F} = \vec{F}_2$ and $\Omega = 0$. The uniqueness of this splitting can be proved in a way similar to that given in Ref. [1].

As a final remark, the corresponding definitions and theorem of splitting for a stationary flow problem can be deduced from the above results by applying a suitable Galilean transformation.

III. Fundamental Solution of One-Dimensional Longitudinal Wave. Inasmuch as the transversal wave is not influenced by compressibility and heat conduction, it would be of interest to single out the longitudinal component by considering some simple cases in order to show the additional heat conducting effect. As a simple example we shall first consider the one-dimensional nonstationary flow, which is itself longitudinal. The fluid motion with spherical symmetry, another example of a flow with a single longitudinal component, will be touched upon in the next section. From these results one may obtain valuable information about the higher dimensional longitudinal waves which will occur as components in more complicated flow phenomena.

Suppose that $\vec{F}$ is in the x-direction, equal to X, and that both X and Ω are given as functions of x and t for all x and $t > 0$ and are zero for $t \leqq 0$. Then the resulting flow, if initially at rest, will also depend only on x and t; the velocity then has only one component $u(x, t)$. In this case the linearized equations (5)–(8) become:

$$p = s + \theta, \tag{21}$$

$$s_t + u_x = 0, \tag{22}$$

$$u_t - (4\nu/3)u_{xx} = -(c^2/\gamma)p_x + \mathrm{X}(x, t), \tag{23}$$

$$\theta_t - \kappa\theta_{xx} = (\gamma - 1)s_t + \Omega(x, t). \tag{24}$$

When the Prandtl number is assumed to be $\frac{3}{4}$, as discussed previously, the system (21)–(24) can be further simplified by elimination to yield:

$$L_1 L_2(\theta) = -\frac{1}{\kappa}\left[\left(1 + \frac{\kappa}{c^2}\frac{\partial}{\partial t}\right)\frac{\partial^2}{\partial x^2} - \frac{\gamma}{c^2}\frac{\partial^2}{\partial t^2}\right]\Omega - \frac{\gamma(\gamma-1)}{\kappa c^2}\,\mathrm{X}_{xt}, \tag{25}$$

$$L_1 L_2(u) = -\frac{1}{\kappa}\,\Omega_{xt} - \frac{\gamma}{c^2}\left(\frac{\partial^2}{\partial x^2} - \frac{1}{\kappa}\frac{\partial}{\partial t}\right)\mathrm{X}_t, \tag{26}$$

$$L_1 L_2(s) = \frac{1}{\kappa}\,\Omega_{xx} + \frac{\gamma}{c^2}\left(\frac{\partial^2}{\partial x^2} - \frac{1}{\kappa}\frac{\partial}{\partial t}\right)\mathrm{X}_x, \tag{27}$$

$$L_1(p) = -c^{-2}\Omega_t + \gamma c^{-2}\mathrm{X}_x \tag{28}$$

where L_1 and L_2 are operators:

$$L_1 \equiv \left(1 + \frac{\kappa}{c^2}\frac{\partial}{\partial t}\right)\frac{\partial^2}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}, \qquad L_2 \equiv \frac{\partial^2}{\partial x^2} - \frac{\gamma}{\kappa}\frac{\partial}{\partial t}. \tag{29}$$

The operator L_1 is the one discussed following Eq. (12) (cf. also Ref. 1, App. A), while L_2 is a typical diffusion operator. The differential equations for θ, u and s are two orders higher than that for p.

Two fundamental solutions will be defined as follows. They show the response of the system (21)–(24) to concentrated addition of a unit amount of heat and concentrated action of a unit force respectively. More precisely, if

$$\mathrm{X} = 0 \quad \text{and} \quad \Omega = \delta(x)\delta(t) \qquad (\delta\text{-Dirac delta function}), \tag{30a}$$

the singular solution of the system, vanishing at infinity, is the fundamental solution due to heat addition, and will be denoted by

$$u = U_\Omega(x, t), \qquad \theta = \Gamma_\Omega(x, t), \qquad s = S_\Omega(x, t), \qquad p = \Pi_\Omega(x, t). \tag{30b}$$

On the other hand, if

$$\mathrm{X} = \delta(x)\delta(t) \quad \text{and} \quad \Omega = 0, \tag{31a}$$

the solution is called the fundamental solution due to external force:

$$u = U_\mathrm{X}(x, t), \qquad \theta = \Gamma_\mathrm{X}(x, t), \qquad s = S_\mathrm{X}(x, t), \qquad p = \Pi_\mathrm{X}(x, t). \tag{31b}$$

For arbitrary Ω and X, the solution under the same zero initial and boundary conditions can be represented by

$$\theta(x, t) = \int_{-\infty}^{\infty} d\xi \int_0^t \left[\Gamma_\Omega(x - \xi, t - \tau)\Omega(\xi, \tau) + \Gamma_\mathrm{X}(x - \xi, t - \tau)\mathrm{X}(\xi, \tau)\right] d\tau \tag{32}$$

together with corresponding formulas for u, p and s. This result follows from the linearity and homogeneity of Eqs. (21)–(24) in space and time. This property also enables one to build up the solution of initial and boundary value problems by using these fundamental solutions[3]. Thus, fundamental solutions represent typical solutions and illustrate many features of the system.

An integral representation of the fundamental solutions can be obtained by taking the Laplace transform of the system with respect to t. Let the Laplace

transform of f be $\bar{f}$:

$$\bar{f}(x, \sigma) = \int_0^\infty e^{-\sigma t} f(x, t)\, dt. \tag{33}$$

The complex inversion formula is then

$$f(x, t) = \frac{1}{2\pi i} \int_{b-i\infty}^{b+i\infty} e^{\sigma t} \bar{f}(x, \sigma)\, d\sigma \tag{34}$$

where the path of integration is parallel to the imaginary σ-axis and to the right of all the singularities of $\bar{f}(x, \sigma)$. Application of the transformation (33) to Eqs. (25)–(28) with zero initial conditions yields:

$$\left(\frac{d^2}{dx^2} - \lambda_1\right)\left(\frac{d^2}{dx^2} - \lambda_2\right)\bar{\theta} = -\frac{1}{\kappa}\left(\frac{d^2}{dx^2} - \gamma\lambda_1\right)\bar{\Omega} - \frac{\gamma(\gamma - 1)\lambda_1}{\kappa\sigma}\overline{X}_x \equiv -M(x, \sigma), \tag{35}$$

$$\left(\frac{d^2}{dx^2} - \lambda_1\right)\bar{p} = -\frac{\lambda_1}{\sigma}\bar{\Omega} + \frac{\gamma\lambda_1}{\sigma^2}\overline{X}_x \equiv -N(x, \sigma), \tag{36}$$

where

$$\lambda_1 = \sigma^2(\kappa\sigma + c^2)^{-1}, \qquad \lambda_2 = \gamma\sigma/\kappa. \tag{37}$$

There are similar equations for $\bar{u}$ and $\bar{s}$. By using the convolution theorem, Eq. (32) is transformed into

$$\bar{\theta}(x, \sigma) = \int_{-\infty}^{\infty} [\bar{\Gamma}_\Omega(x - \xi, \sigma)\bar{\Omega}(\xi, \sigma) + \bar{\Gamma}_X(x - \xi, \sigma)\overline{X}(\xi, \sigma)]\, d\xi. \tag{38}$$

The Green's function $G^{(1)}$ of Eq. (36), defined by

$$\bar{p}(x, \sigma) = \int_{-\infty}^{\infty} G^{(1)}(x - \xi, \sigma)N(\xi, \sigma)\, d\xi \tag{39a}$$

is known to be (for example, Ref. 9, p. 376)

$$G^{(1)}(x, \sigma) = \tfrac{1}{2}\lambda_1^{-\frac{1}{2}} \exp(-\lambda_1^{\frac{1}{2}} |\, x\,|). \tag{39b}$$

If we denote the Green's function of Eq. (35) by $G^{(2)}$, where

$$\bar{\theta}(x, \sigma) = \int_{-\infty}^{\infty} G^{(2)}(x - \xi, \sigma)M(\xi, \sigma)\, d\xi, \tag{40a}$$

we can show that the factoring of the operator in Eq. (35) enables us to represent $G^{(2)}$ as a sum of functions of the type $G^{(1)}$ (cf. Ref. 3, App. A):

$$G^{(2)}(x, \sigma) = \tfrac{1}{2}(\lambda_1 - \lambda_2)^{-1}\{\lambda_1^{-\frac{1}{2}} \exp(-\lambda_1^{\frac{1}{2}} |\, x\,|) - \lambda_2^{-\frac{1}{2}} \exp(-\lambda_2^{\frac{1}{2}} |\, x\,|)\}. \tag{40b}$$

By substituting the values of M and N into (40a) and (39a), then integrating by parts the terms with $G^{(2)}\bar{\Omega}_{\xi\xi}$ and $G^{(2)}\overline{X}_\xi$, and comparing the result with Eq. (38), the solutions $\bar{\Gamma}_\Omega$, $\bar{\Gamma}_X$ etc. can be related to $G^{(1)}$ and $G^{(2)}$. Finally, if we apply

the inversion formula (34) to these expressions for $\bar{\Gamma}_0$, $\bar{\Gamma}_X$ etc., and introduce the nondimensional quantities:

$$z = \kappa\sigma/c^2, \qquad t' = c^2t/\kappa, \qquad x' = c|x|/\kappa, \tag{41}$$

we obtain the following integral representations for these fundamental solutions:

$$\Gamma_0 = \frac{c}{4\kappa\pi i}\int_C e^{zt'}\left(z + \frac{\gamma}{\gamma - 1}\right)^{-1} \cdot \left\{(z+1)^{\frac{1}{2}}e^{-z(z+1)^{-\frac{1}{2}}x'} + \frac{(\gamma/z)^{\frac{1}{2}}}{\gamma - 1}\, e^{-(\gamma z)^{\frac{1}{2}}x'}\right\} dz, \tag{42a}$$

$$U_0 = \frac{c^2 \operatorname{sgn} x}{4(\gamma - 1)\kappa\pi i}\int_C e^{zt'}\left(z + \frac{\gamma}{\gamma - 1}\right)^{-1}\{e^{-z(z+1)^{-\frac{1}{2}}x'} - e^{-(\gamma z)^{\frac{1}{2}}x'}\}\, dz, \tag{42b}$$

$$\Pi_0 = \frac{c}{4\kappa\pi i}\int_C (z+1)^{-\frac{1}{2}} \exp\{zt' - z(z+1)^{-\frac{1}{2}}x'\}\, dz, \tag{42c}$$

$$\Gamma_X = \frac{\gamma(\gamma - 1)}{c^2}\, U_0, \tag{43a}$$

$$U_X = \frac{c}{4\kappa\pi i}\int_C e^{zt'}\left(z + \frac{\gamma}{\gamma - 1}\right)^{-1} \cdot \left\{\frac{\gamma(z+1)^{-\frac{1}{2}}}{\gamma - 1}\, e^{-z(z+1)^{-\frac{1}{2}}x'} + (\gamma z)^{\frac{1}{2}}e^{-(\gamma z)^{\frac{1}{2}}x'}\right\} dz, \tag{43b}$$

$$\Pi_X = \frac{\gamma \operatorname{sgn} x}{4\kappa\pi i}\int_C (z+1)^{-1} \exp\{zt' - z(z+1)^{-\frac{1}{2}}x'\}\, dz \tag{43c}$$

and the density is simply given by $S = \Pi - \Gamma$. The contour C is the same as that given in Eq. (34).

For general values of x' and t' it does not seem possible that these contour integrals can be evaluated in closed form. However, for t' either large or small these integrals can be transformed into appropriate forms to which various methods for asymptotic expansion can readily be applied. The detailed calculation is shown in the Appendix, and the final results can be summarized as follows:

(a) For large values of $t' = c^2t/\kappa$, all the following formulas are accurate within a factor $[1 + 0(1/t')]$:

$$\Gamma_0(x,t) = \frac{1}{\gamma}\left(\frac{\gamma}{4\pi\kappa t}\right)^{\frac{1}{2}} \exp\left[-\frac{\gamma x^2}{4\kappa t}\right] + \frac{\gamma - 1}{2\gamma}\,\mathfrak{F}(x,t), \tag{44a}$$

$$S_0(x,t) = -\frac{1}{\gamma}\left(\frac{\gamma}{4\pi\kappa t}\right)^{\frac{1}{2}} \exp\left[-\frac{\gamma x^2}{4\kappa t}\right] + \frac{1}{2\gamma}\,\mathfrak{F}(x,t), \tag{44b}$$

$$U_0(x,t) = \frac{c}{2\gamma}(\operatorname{sgn} x)\mathfrak{F}(x,t), \qquad \Pi_0(x,t) = \tfrac{1}{2}\mathfrak{F}(x,t), \tag{44c}$$

$$\Gamma_X(x,t) = \frac{\gamma - 1}{2c}(\operatorname{sgn} x)\mathfrak{F}(x,t), \qquad U_X(x,t) = \tfrac{1}{2}\mathfrak{F}(x,t), \tag{45a}$$

$$S_{\mathfrak{X}}(x,t) = \frac{1}{2c}(\operatorname{sgn} x)\mathfrak{F}(x,t), \qquad \Pi_{\mathfrak{X}}(x,t) = \frac{\gamma}{2c}(\operatorname{sgn} x)\mathfrak{F}(x,t), \tag{45b}$$

where $\mathfrak{F}(x, t)$ denotes the function

$$\mathfrak{F}(x,t) = \left(\frac{1}{2\pi\kappa t}\right)^{\frac{1}{2}}\left\{\exp\left[-\frac{(x-ct)^2}{2\kappa t}\right] + \exp\left[-\frac{(x+ct)^2}{2\kappa t}\right]\right\}. \tag{46}$$

This function exhibits the diffusion of concentrated disturbances about the two opposite running sound wave fronts $x = \pm ct$.

(b) For small values of $t' = c^2t/\kappa$, all the following results are accurate within a factor $[1 + 0(t')]$:

$$\Gamma_{\mathfrak{Q}}(x,t) = \Pi_{\mathfrak{Q}}(x,t) = \left(\frac{1}{4\pi\kappa t}\right)^{\frac{1}{2}}\exp\left[-\frac{x^2}{4\kappa t}\right], \tag{47a}$$

$$U_{\mathfrak{Q}}(x,t) = \frac{c^2 \operatorname{sgn} x}{2\kappa(\gamma-1)}\left\{\operatorname{erfc}\left(\frac{x^2}{4\kappa t}\right)^{\frac{1}{2}} - \operatorname{erfc}\left(\frac{\gamma x^2}{4\kappa t}\right)^{\frac{1}{2}}\right\}, \tag{47b}$$

$$S_{\mathfrak{Q}}(x,t) = \frac{c^2}{\kappa(\gamma-1)}\left(\frac{t}{\pi\kappa}\right)^{\frac{1}{2}}\left\{\exp\left[-\frac{x^2}{4\kappa t}\right] - \gamma^{\frac{1}{2}}\exp\left[-\frac{\gamma x^2}{4\kappa t}\right]\right\}, \tag{47c}$$

$$\Gamma_{\mathfrak{X}}(x,t) = \frac{\gamma(\gamma-1)}{c^2}U_{\mathfrak{Q}}, \tag{48a}$$

$$U_{\mathfrak{X}}(x,t) = \left(\frac{\gamma}{4\pi\kappa t}\right)^{\frac{1}{2}}\exp\left[-\frac{\gamma x^2}{4\kappa t}\right], \tag{48b}$$

$$S_{\mathfrak{X}}(x,t) = \frac{\gamma}{2\kappa}(\operatorname{sgn} x)\operatorname{erfc}\left(\frac{\gamma x^2}{4\kappa t}\right)^{\frac{1}{2}}. \tag{48c}$$

Thus for t' small, diffusion of all flow quantities is the main process.

IV. Discussion of Results. Several interesting points may be noted from the nature of the flow field described by the above asymptotic formulas. Let us consider first the solutions due to heat addition Ω. The over-all conservation of heat energy can be obtained by integrating Eqs. (8) and (6) over the domain $-\infty < x < \infty$ and $0 \leqq \tau \leqq t$ to yield

$$\int_{-\infty}^{\infty}\theta(x,t)\,dx = \int_{-\infty}^{\infty}dx\int_{0}^{t}\Omega(x,\tau)\,d\tau \tag{49}$$

which, for the case $\Omega(x, t) = \delta(x)\delta(t)$, reduces to

$$\int_{-\infty}^{\infty}\Gamma_{\mathfrak{Q}}(x,t)\,dx = 1 \qquad \text{for all } t > 0. \tag{50}$$

This relation is satisfied by the above asymptotic formulas for both small and large t'. More specifically, for small t' all the added heat diffuses about the origin with diffusivity κ. However, for t' large, only $1/\gamma$ of the originally added heat is left to diffuse about the origin with diffusivity κ/γ (cf. Eq. 44a), while the rest $(1 - 1/\gamma)$ of the total heat diffuses about the propagating sound wave fronts

$x = \pm ct$ with an effective diffusivity $\kappa/2$. The time history of this interesting transition can be seen rather clearly by integrating the two terms in Eq. (42a) with respect to x. Write Eq. (42a) as: $\Gamma_0(x, t) = \Gamma_{0_1} + \Gamma_{0_2}$ with

$$\Gamma_{0_1} = \frac{c}{4\pi\kappa i}\int_C \left(z + \frac{\gamma}{\gamma - 1}\right)^{-1} (z + 1)^{\frac{1}{2}} \exp\{zt' - z(z + 1)^{-\frac{1}{2}}x'\}\, dz, \quad (51a)$$

$$\Gamma_{0_2} = \frac{c\gamma^{\frac{1}{2}}}{4\pi(\gamma - 1)\kappa i}\int_C \left(z + \frac{\gamma}{\gamma - 1}\right)^{-1} z^{-\frac{1}{2}} \exp\{zt' - (\gamma z)^{\frac{1}{2}}x'\}\, dz, \quad (51b)$$

where x' and t' are given by Eq. (41). Then for $t > 0$,

$$H_1 = \int_{-\infty}^{\infty} \Gamma_{0_1}\, dx = \frac{1}{2\pi i}\int_C \frac{(z + 1)\exp(zt')}{z[z + \gamma/(\gamma - 1)]}\, dz = \frac{\gamma - 1}{\gamma} + \frac{1}{\gamma}\exp\left[-\frac{\gamma c^2 t}{(\gamma - 1)\kappa}\right], \quad (52a)$$

$$H_2 = \int_{-\infty}^{\infty} \Gamma_{0_2}\, dx = \frac{(\gamma - 1)^{-1}}{2\pi i}\int_C \frac{\exp(zt')}{z[z + \gamma/(\gamma - 1)]}\, dz = \frac{1}{\gamma} - \frac{1}{\gamma}\exp\left[-\frac{\gamma c^2 t}{(\gamma - 1)\kappa}\right] \quad (52b)$$

so that their sum is still unity for all $t > 0$. Now as $t \to 0$, $H_1 \to 1$ and $H_2 \to 0$. However, as $t \to \infty$, $H_1 \to (\gamma - 1)/\gamma$ and $H_2 \to 1/\gamma$. Obviously Γ_{0_2} is significant only for large t while Γ_{0_1} (due to the operator L_1) changes the role of diffusion about the origin for small t to that of diffusion about the propagating sound waves for t large. The effective time duration of such transition (see Eq. 52) is apparently of the order of $(\gamma - 1)\kappa/(\gamma c^2) = \tau_0$, say. It then becomes definite for what we have meant by "t large" and "t small". These two stages may be defined by $t \gg \tau_0$ and $t < \tau_0$. Thus we see from this analysis that in linear theory the sound wave requires a time, approximately τ_0, to be established; before that instant simple diffusion is the main process. The value of τ_0 is in general very small. For atmospheric air it is of the order of 10^{-8} sec.

For other flow quantities one can see from Eqs. (47) that for $t < \tau_0$ significant changes, which take place only near the origin, are mainly in temperature and pressure, the velocity variation is negligible while the density field builds up at a rate proportional to $t^{\frac{1}{2}}$. For $t \gg \tau_0$ the sound wave is then formed, around which all variables share equal importance. More precisely, near the sound waves the thermodynamic variables satisfy the linearized isentropic relations

$$p = \gamma s = \frac{\gamma}{\gamma - 1}\theta \quad (53)$$

and the velocity-density relation is actually the Riemann invariant:

$$u = \pm cs \quad \text{near} \quad x = \pm ct. \quad (54)$$

Equations (53) and (54) are the familiar results of the inviscid acoustic theory. It is of particular interest to note that near the origin the thermodynamic process

changes from mostly isosteric (constant s) for $t < \tau_0$ to isobaric (constant p) for $t > \tau_0$. This transition, through which the pressure waves are formed, is made at the cost of spending all the pressure accumulated at first near the origin by conveying it to the wave fronts. It thus leaves a negative density perturbation to balance the temperature field near the origin (see Eqs. 44a, b).

This result can be used to verify (thermodynamically) the final heat distribution*. Since the process is isosteric for $t < \tau_0$, an element of instantaneously added heat ΔQ will result in a temperature increment $\Delta\theta$, and hence a pressure increment $(\Delta p)_1$, according to

$$\Delta Q = c_v T_0 \Delta\theta = c_v T_0 (\Delta p)_1 . \tag{55}$$

During the transition right after the heat addition, Δp decreases and a density perturbation Δs builds up. Hence a certain amount of work is being done by the fluid near the origin. The process being adiabatic after the heat addition is completed, this work must be equal to the decrease in internal energy:

$$\Delta E = -\Delta W = -P\Delta(\rho^{-1}) \cong P_0 \Delta s/\rho_0 = RT_0(1/\gamma)(\Delta p)_2 . \tag{56}$$

But our previous result shows that the final pressure perturbation vanishes after the transition (that is, $t \gg \tau_0$), we must have

$$(\Delta p)_1 + (\Delta p)_2 = 0, \tag{57}$$

which then leads to

$$\Delta E = -\frac{\gamma - 1}{\gamma} \Delta Q. \tag{58}$$

This decrease in internal energy of the fluid near the origin must become the heat energy carried away by the propagating sound waves. The rest of the added heat, $(1/\gamma)\Delta Q$, is then left to diffuse about the origin.

The solutions for X have very similar features in general with the important exception that the thermodynamic process near the origin changes from isothermal for $t < \tau_0$ to isobaric for $t > \tau_0$.

The linearized flow with spherical symmetry can be worked out in a similar way. For somplicity we shall consider only the flow generated by the heat addition $\Omega(r, t) = (4\pi r^2)^{-1}\delta(r - r_0)\delta(t)$ where r is the radial distance from the origin and $r = r_0$ is the spherical shell at which the heat is added. Only the results will be presented here: the following formulas are asymptotic for large values of $c^2 t/\kappa$, with a factor of error $[1 + 0(\kappa/c^2 t)]$.

$$\theta_\Omega = \frac{1}{4\pi r_0 r}\left[\frac{1}{\gamma}\mathcal{G}(r, t) + \frac{\gamma - 1}{2\gamma}\mathcal{F}(r, t)\right], \qquad p_\Omega = \frac{1}{8\pi r_0 r}\mathcal{F}(r, t),$$

$$u_\Omega = \frac{c}{8\pi\gamma r_0 r}\left[\mathcal{K}(r, t) + \frac{1}{2r}\mathcal{J}(r, t) - \frac{2\kappa}{c\gamma r}\mathcal{G}(r, t)\right]$$

and $s_\Omega = p_\Omega - \theta_\Omega$, where

$$\mathcal{F}(r, t) = \left(\frac{1}{2\pi\kappa t}\right)^{\frac{1}{2}}\left[e^{-((r-r_0-ct)^2/2\kappa t)} + e^{-((r-r_0+ct)^2/2\kappa t)} - e^{-((r+r_0-ct)^2/2\kappa t)}\right],$$

* This thermodynamic verification of the final result was shown to the author by Professor H. S. Tsien during their discussion of this paper.

$$\mathcal{G}(r,t) = \left(\frac{\gamma}{4\pi\kappa t}\right)^{\frac{1}{2}} [e^{-(\gamma(r-r_0)^2/4\kappa t)} - e^{-(\gamma(r+r_0)^2/4\kappa t)}],$$

$$\mathcal{K}(r,t) = \frac{\operatorname{sgn}(r-r_0)}{(2\pi\kappa t)^{\frac{1}{2}}} [e^{-((r-r_0-ct)^2/2\kappa t)} + e^{-((r-r_0+ct)^2/2\kappa t)}] - \left(\frac{1}{2\pi\kappa t}\right)^{\frac{1}{2}} e^{-((r-r_0-ct)^2/2\kappa t)},$$

$$\mathcal{I}(r,t) = \operatorname{erfc}\left[\frac{(r-r_0-ct)^2}{2\kappa t}\right]^{\frac{1}{2}} + \operatorname{erfc}\left[\frac{(r-r_0+ct)^2}{2\kappa t}\right]^{\frac{1}{2}} - \operatorname{erfc}\left[\frac{(r+r_0-ct)^2}{2\kappa t}\right]^{\frac{1}{2}}.$$

And the asymptotic solutions for small values of c^2t/κ are

$$\theta_0 = p_0 = \frac{1}{4\pi r_0 r}\left(\frac{1}{4\pi\kappa t}\right)^{\frac{1}{2}} [e^{-((r-r_0)^2/4\kappa t)} - e^{-(r+r_0)^2/4\kappa t}];$$

and the variations of s_0 and u_0 are of $0(t'^{\frac{1}{2}})$ and are less important.

The limiting values of these solutions as (i) $r_0 \to 0$ and (ii) $r_0 \to \infty$ are easy to deduce. The second case reduces to the one-dimensional problem if $4\pi r_0^2\Omega$ is regarded as the corresponding one-dimensional heat strength.

For flows with a large constant velocity at infinity the equations of motion can be deduced from our present case by using a Galilean transformation. Such a stationary flow for the two-dimensional case is worked out in Reference [10] with an associated boundary value problem.

Appendix. Asymptotic Expansions of the Fundamental Solution. The integral representation for $\Gamma_0 = \Gamma_{0_1} + \Gamma_{0_2}$ as given by Eq. (51) will be evaluated asymptotically for t' large or small, and is typical for all the integrals. We first apply the convolution theorem to Eq. (51b), regarding the integrand as the product of $[z + \gamma/(\gamma - 1)]^{-1}$ and $z^{-\frac{1}{2}} \exp[-(\gamma z)^{\frac{1}{2}}x']$. Then we have

$$\Gamma_{0_2} = \frac{ac}{2\kappa}\left(\frac{t'}{\gamma\pi}\right)^{\frac{1}{2}} e^{-at'} \int_0^1 \eta^{-\frac{1}{2}} \exp\left\{at'\eta - \left(\frac{\gamma x'^2}{4t'}\right)\frac{1}{\eta}\right\} d\eta, \tag{A-1}$$

where $a = \gamma/(\gamma - 1)$. This integral representation is convenient for approximating for either small or large values of t', and is also suitable for numerical computation with given x' and t'. In the integrand there are two similarity parameters, x^2/t and t. For small t' we can expand $\exp(at'\eta)$ into a series and then integrate termwise. The final result is

$$\Gamma_{0_2} = \frac{c^2}{(\gamma - 1)\kappa}\left(\frac{\gamma t}{\pi\kappa}\right)^{\frac{1}{2}} \left\{\exp\left(-\frac{\gamma x^2}{4\kappa t}\right) - \left(\frac{\gamma\pi x^2}{4\kappa t}\right)^{\frac{1}{2}} \operatorname{erfc}\left(\frac{\gamma x^2}{4\kappa t}\right)^{\frac{1}{2}}\right\} [1 + 0(t')]. \tag{A-2}$$

When t' is large one can integrate Eq. (A-1) by parts to obtain the asymptotic formula:

$$\Gamma_{0_2} \cong \frac{1}{\gamma}\left(\frac{\gamma}{4\pi\kappa t}\right)^{\frac{1}{2}} \exp\left(-\frac{\gamma x^2}{4\kappa t}\right)\left[1 + \frac{\gamma - 1}{2\gamma}\frac{\kappa}{c^2 t}\left(1 - \frac{\gamma x^2}{4\kappa t}\right) + 0\left(\frac{\kappa}{c^2 t}\right)^2\right] \tag{A-3}$$

which is essentially diffusive in nature.

In Eq. (51a) we first remove the branch point at $z = -1$ by the conformal transformation $\zeta^2 = z + 1$. Then we obtain

$$\Gamma_{0_1} = \frac{c}{2\kappa\pi i}\int_{C_1} \zeta^3\left(\zeta^2 + \frac{1}{\gamma - 1}\right)^{-1} \exp\left\{t'(\zeta^2 - 1)\left(1 - \frac{\omega}{\zeta}\right)\right\} d\zeta \tag{A-4}$$

where $\omega = x'/t' = |x|/(ct)$ and is regarded as a positive real parameter; the path C_1 may be deformed in such a way that it starts from $\infty e^{-i\alpha}$, crosses the positive real axis and ends up at $\infty e^{i\alpha}$ with $\pi/4 < \alpha < 3\pi/4$.

When t' is large, the method of steepest descent[11] may be applied to evaluate this integral asymptotically; the result will contain an undetermined parameter ω. Write $f(\zeta) = (\zeta^2 - 1)(1 - \omega/\zeta)$, then the saddle points are the zeros of $f'(\zeta)$, i.e., the roots of $2\zeta^3 - \omega\zeta^2 - \omega = 0$. This equation has one positive real root, $\zeta = \zeta_0(\omega)$, and two complex conjugate roots with a negative real part. Here $\zeta_0(\omega)$ is the only relevant saddle point. Then according to the method, the asymptotic value of the above integral will contain the term $\exp[t'f(\zeta_0)]$. In this case $f(\zeta_0(\omega), \omega) \leqq 0$ for $\omega \geqq 0$, and $f(\zeta_0)$ reaches its maximum value (zero) at

$$\frac{d}{d\omega} f(\zeta_0(\omega), \omega) = f'(\zeta_0)\frac{d\zeta_0}{d\omega} + \frac{\partial f}{\partial \omega} = -(\zeta_0^2 - 1)/\zeta_0 = 0,$$

or $\zeta_0 = 1$ ($\zeta_0 = -1$ is not considered since the contour C_1 must cross the positive real axis). When $\zeta_0 = 1$, the condition $f'(\zeta_0(\omega)) = 0$ then yields $\omega = 1$ ($\sim x' = t'$, or $x = \pm ct$). This result indicates that for large t' (corresponding to vanishing κ) the disturbances represented by Γ_{Ω_1} will concentrate about the sonic wave fronts $x = \pm ct$, a fact already predicted by the characteristics.Now we write the exponential function in (A-4) as

$$\exp\{t'(\zeta - 1)^2(1 + \zeta^{-1})\} \cdot \exp\{(t' - x')(\zeta^2 - 1)/\zeta\}$$

and assume that $|t' - x'|$ is bounded as $t' \to \infty$. Then the path of steepest descent C_1 passes through the saddle point $\zeta = 1$ with

$$\operatorname{Im}\{(\zeta - 1)^2(1 + \zeta^{-1})\} = 0$$

on C_1. Since most of the contribution to the integral comes from the neighborhood of $\zeta = 1$, C_1 may be approximated near $\zeta = 1$ by

$$\zeta = \xi + i\eta = (1 - \eta^2/4) + i\eta$$

together with the same approximation for the functions in the integrand. Consequently we obtain

$$\Gamma_{\Omega_1} \cong \frac{(\gamma - 1)c}{2\gamma\kappa\pi} \int_{-\infty}^{\infty} e^{-2t'\eta^2 + 2i(t' - x')\eta}\, d\eta = \frac{\gamma - 1}{2\gamma} \mathfrak{F}(x, t)\left[1 + O\left(\frac{1}{t'}\right)\right] \quad \text{(A-5)}$$

where $\mathfrak{F}(x, t)$ is given by Eq. (46).

For small values of t' the attenuation of physical quantities should be governed by the parameter x^2/t, as can be seen from (A-1). Hence we first apply to Eq. (A-4) the transformation $w = (t')^{\frac{1}{2}}\zeta$, and then approximate the integral by assuming that t' is small but x'^2/t' is of lower order than t'. Finally we obtain

$$\Gamma_{\Omega_1} \cong \frac{ct'^{-\frac{1}{2}}}{2\kappa\pi i} \int_{-i\infty}^{i\infty} \exp\{w^3 - x'w/t'^{\frac{1}{2}}\}\, dw[1 + O(t')]$$

$$\cong \left(\frac{1}{4\pi\kappa t}\right)^{\frac{1}{2}} \exp\left(-\frac{x^2}{4\kappa t}\right)\left[1 + O\left(\frac{c^2 t}{\kappa}\right)\right]. \quad \text{(A-6)}$$

REFERENCES

1. Lagerstrom, P. A., Cole, J. D. and Trilling, L.: Problems in the Theory of Viscous Compressible Fluids. Monograph, California Institute of Technology, March 1949.
2. Cole, J. D. and Wu, T. Y.: Heat Conduction in a Compressible Fluid. J. Appl. Mech. **19**, p. 209 (1952).
3. Wu, T. Y.: On Problems of Heat Conduction in a Compressible Fluid. Ph.D. Thesis, California Institute of Technology, 1952.
4. Kovasznay, L. S. G.: Turbulence in Supersonic Flow. J. Aero. Sci., **20**, p. 657 (1953).
5. Howarth, L. (editor): Modern Developments in Fluid Dynamics. Vol. I, Oxford Press, 1953.
6. Lamb, H.: Hydrodynamics. 6th Edition, Dover Publications, New York, 1945.
7. Oseen, C. W.: Hydrodynamik. 1927.
8. Kemke, E.: Differentialgleichungen reeller Funktionen. Chelsea Pub. Co., New York, 1947.
9. Courant, R. and Hilbert, D.: Methods of Mathematical Physics. Vol. 1, p. 376, Interscience Publisher, New York, 1953.
10. Wu, T. Y. and Cole, J. D.: Anemometry of a Heated Flat Plate. Heat Transfer and Fluid Mechanics Institute, University of California at Los Angeles, 1952.
11. Jeffreys, H. and Jeffreys, B. S.: Methods of Mathematical Physics. Cambridge Press, 1950.

Guggenheim Aeronautical Laboratory
California Institute of Technology

(Received July 10, 1955)

VIBRATION AND STABILITY PROBLEMS OF BEAMS TREATED BY MATRICES*

By K. Marguerre

1. Introduction. In vibration problems, matrix calculus[1] is known in connection with the method of influence-coefficients.[2] The method of Myklestad[3] is apparently more effective than the influence coefficient method for determining the higher modes of vibration of a shaft of variable cross-section with intermediate supports (critical speed of machinery). Myklestad's method, however, can also be interpreted as a matrix method,[4] a fact which greatly facilitates understanding and handling of the procedure. We therefore give a short survey of the matrix method for the differential equation of a vibrating beam, and for the related problem of the column.[5] It turns out that the equation for the eigenvalue appears always as the condition that a second order subdeterminant of the governing matrix must vanish (in the influence-coefficient method the determinant is of the order of the number of degrees of freedom). Amazingly, systematic difficulties arise in the especially interesting cases of intermediate rigid supports or ideal pin connections. Even if the supports or the hinges are elastic and therefore can be taken into account by a transition matrix, another very inconvenient thing can occur: the coefficients of the equation for the eigenvalues appear as differences of large numbers. Both difficulties can be overcome by introducing the notion of a kind of associated matrix, called the Δ-matrix, for which simple rules of calculation can be given.

2. Vibration and Buckling of Beams Computed by Matrix Methods. The differential equations for the vibration and for the buckling of a uniform beam (μ mass per unit length, P axial load) are well known:

$$\begin{aligned} EIw^{IV} - \omega^2\mu w &= 0 & \qquad EIw^{IV} + Pw'' &= 0 \\ \text{or} \quad w^{IV} - \frac{\lambda^4}{l^4} w &= 0 & \qquad \text{or} \quad w^{IV} + \frac{\lambda^2}{l^2} w'' &= 0 \\ \text{with} \quad \lambda^4 &= \frac{\omega^2\mu l^4}{EI} & \qquad \text{with} \quad \lambda^2 &= \frac{Pl^2}{EI} \end{aligned} \tag{1}$$

w denoting the deflection. The solutions of these equations

* This paper was prepared under the sponsorship of the Office of Scientific Research of the United States Airforce while the author was a Visiting Professor at Washington University, St. Louis, Missouri.

[1] Further details will be found in two Dissertations: H. Fuhrke, Bestimmung von Stabwerkschwingungen mit Hilfe des Matrizencalcüls, Ing. Arch. 23, p. 329 (1955); W. Schnell, Krafteinleitung in versteifte Cylinderschalen, Zeitschrift f. Flugwissenschaften 1955. Also: W. Schnell, ZAMM 35, p. 296 (1955).

[2] Thomson, Mechanical Vibrations, 2nd Ed., 1953, p. 169.

[3] Myklestad, J. Aero. Sc., 1944, p. 153.

[4] Thomson, J. Appl. Mech., 1950, p. 337.

[5] The procedure so far is not "new", but we do not know what paper could be quoted. Thomson, e.g., J. Aero. Sc. 1949, p. 623, deals only with the special case that no transverse force occurs in the axially compressed column.

$$w = A\cos\lambda\frac{x}{l} + B\sin\lambda\frac{x}{l} + C\,\mathrm{Cosh}\,\lambda\frac{x}{l} + D\,\mathrm{Sinh}\,\lambda\frac{x}{l}$$

$$w = A\cos\lambda\frac{x}{l} + B\sin\lambda\frac{x}{l} + C\frac{x}{l} + D$$

can be arranged in a manner such that instead of the four constants $A \cdots D$ (which have no obvious mechanical meaning), the deflection, slope, moment and shearing force at the point $x = 0$ appear as constants of integration. (For a reason which will readily be understood from Fig. 2 below, we choose $x = 0$ at the *right*-hand end of the beam and count x towards the left.) For the vibrating beam the 4 mechanically interesting quantities are (see Fig. 1)

$$w, \quad \psi = w', \quad M = EI\psi', \quad \text{and} \quad Q = M'$$

In the case of the column, instead of $Q = M'$ we have to use the combination $V = M' + Pw'$, because in the presence of a large axial force P it is necessary to state clearly that at a "free" end not the shearing force acting on the inclined surface disappears but rather the vertical force V (perpendicular to P).[6]

The calculation which replaces $A \cdots D$ by $w(0) = w_0, \psi_0, M_0, V_0$ is elementary and leads to the following results:

$$w(x) = w_0\,\tfrac{1}{2}\left(\mathrm{Cosh}\,\lambda\frac{x}{l} + \cos\lambda\frac{x}{l}\right) + \psi_0 l\frac{1}{2\lambda}\left(\mathrm{Sinh}\,\lambda\frac{x}{l} + \sin\lambda\frac{x}{l}\right) + M_0\frac{l^2}{EI}\frac{1}{2\lambda^2}\left(\mathrm{Cosh}\,\lambda\frac{x}{l} - \cos\lambda\frac{x}{l}\right) + V_0\frac{l^3}{EI}\frac{1}{2\lambda^3}\left(\mathrm{Sinh}\,\lambda\frac{x}{l} - \sin\lambda\frac{x}{l}\right)$$

$$w(x) = w_0 + \psi_0 l\frac{1}{\lambda}\sin\lambda\frac{x}{l} + M_0\frac{l^2}{EI}\frac{1}{\lambda^2}\left(1 - \cos\lambda\frac{x}{l}\right) + V_0\frac{l^3}{EI}\frac{1}{\lambda^3}\left(\lambda\frac{x}{l} - \sin\lambda\frac{x}{l}\right) \qquad (2)$$

The primary quantity of interest is the eigenvalue, which is found from the condition that the four boundary conditions at $x = 0$, $x = l$ are homogeneous equations for the four constants $w_0 \cdots V_0$. This leads to non-trivial solutions only in the case that a characteristic determinant vanishes. As two of the four constants are immediately known from the conditions at $x = 0$, only two equations need to be written down explicitly and the determinant just mentioned will be of the second order. We introduce the following abbreviations:

$$C_0 = \tfrac{1}{2}(\mathrm{Cosh}\,\lambda + \cos\lambda), \quad S_1 = \frac{1}{2\lambda}(\mathrm{Sinh}\,\lambda + \sin\lambda), \quad C_2 = \frac{1}{2\lambda^2}(\mathrm{Cosh}\,\lambda - \cos\lambda), \quad S_3 = \frac{1}{2\lambda^3}(\mathrm{Sinh}\,\lambda - \sin\lambda)$$

$$C_0 = \cos\lambda \quad S_1 = \frac{1}{\lambda}\sin\lambda \quad C_2 = \frac{1}{\lambda^2}(1 - \cos\lambda), \quad S_3 = \frac{1}{\lambda^3}(\lambda - \sin\lambda) \qquad (3)$$

[6] The simplest proof of the fact that V and not M' has to be taken as the fourth quantity at $x = 0$ is provided by the energy method where V appears as the "generalized force" corresponding to the displacement w.

and we call $w(l)$ etc., shortly w_1 etc. Then the 4 equations connecting the 4 quantities $w_1 \cdots$ with $w_0 \cdots$ can conveniently be written as a matrix equation:

$$\begin{Vmatrix} w \\ l\psi \\ \frac{l^2}{EI} M \\ \frac{l^3}{EI} V \end{Vmatrix}_1 = \begin{pmatrix} C_0 & S_1 & C_2 & S_3 \\ \lambda^4 S_2 & C_0 & S_1 & C_2 \\ \lambda^4 C_2 & \lambda^4 S_3 & C_0 & S_1 \\ \lambda^4 S_1 & \lambda^4 C_2 & \lambda^4 S_3 & C_0 \end{pmatrix} \begin{Vmatrix} w \\ l\psi \\ \frac{l^2}{EI} M \\ \frac{l^3}{EI} V \end{Vmatrix}_0 \quad \left| \quad \begin{Vmatrix} w \\ l\psi \\ \frac{l^2}{EI} M \\ \frac{l^3}{EI} V \end{Vmatrix}_1 = \begin{pmatrix} 1 & S_1 & C_2 & S_3 \\ 0 & C_0 & S_1 & C_2 \\ 0 & -\lambda^2 S_1 & C_0 & S_1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{Vmatrix} w \\ l\psi \\ \frac{l^2}{EI} M \\ \frac{l^3}{EI} V \end{Vmatrix}_0 \right. \tag{4}$$

or

$$\{w\}_1 = \mathfrak{R} \{w\}_0 \quad \left| \quad \{w\}_1 = \mathfrak{K} \{w\}_0 \right. \tag{4'}$$

$\{w\}$ denoting the column-matrix with the four "components" of equal dimension

$$w_1 = w, \qquad w_2 = l\psi, \qquad w_3 = \frac{l^2}{EI} M, \qquad w_4 = \frac{l^3}{EI} V \tag{4''}$$

The two matrices *transforming* $\{w\}_0$ into $\{w\}_1$ are called $\mathfrak{R} = [R_{ik}]$ and $\mathfrak{K} = [K_{ik}]$, the first in honor of Lord Rayleigh who introduced the combinations $\operatorname{Cosh} \lambda x/l + \cos \lambda x/l$ etc. into the solution of Eq. (1).

The matrices $\mathfrak{R}$ and $\mathfrak{K}$ are symmetrical with respect to the cross diagonal ("cross-symmetrical"). Their determinant is 1, and the reciprocal matrices $\mathfrak{R}^{-1}$, $\mathfrak{K}^{-1}$ are identical with $\mathfrak{R}$, $\mathfrak{K}$ themselves, their elements, however, taken with alternating signs. This means mechanically that the transition from the point "1" to the point "0" has the same matrix as that from "0" to "1" except that, obviously, the signs of ψ and V are reversed.

The sense of the matrix notation is that for an n-section beam (see Fig. 2) Eq. (4) in a simple manner can be generalized: If, instead of using l_i and EI_i of every field for the definition of the vector $\{w\}_i$ we introduce two quantities l and EI "of reference" say l_0, EI_0, the condition of continuity of w, ψ, M, V at the points $(1', 2)$, $(2', 3)$ etc. yields the continuity of the "vector" $\{w\}$. If, therefore, with

$$\alpha_i = \frac{EI_i}{EI_0}, \qquad \beta_i = \frac{l_i}{l_0} \tag{5}$$

we write Eq. (4) in the generalized form

$$\{w\}_{i'} = \begin{pmatrix} C_0 & \beta S_1 & \frac{\beta^2}{\alpha} C_2 & \frac{\beta^3}{\alpha} S_3 \\ \frac{1}{\beta} \lambda^4 S_3 & C_0 & \frac{\beta}{\alpha} S_1 & \frac{\beta^2}{\alpha} C_2 \\ \frac{\alpha}{\beta^2} \lambda^4 C_2 & \frac{\alpha}{\beta} \lambda^4 S_3 & C_0 & \beta S_1 \\ \frac{\alpha}{\beta^3} \lambda^4 S_1 & \frac{\alpha}{\beta^2} \lambda^4 C_2 & \frac{1}{\beta} \lambda^4 S_3 & C_0 \end{pmatrix}_i \{w\}_i \quad \left| \quad \{w\}_{i'} = \begin{pmatrix} 1 & \beta S_1 & \frac{\beta^2}{\alpha} C_2 & \frac{\beta^3}{\alpha} S_3 \\ 0 & C_0 & \frac{\beta}{\alpha} S_1 & \frac{\beta^2}{\alpha} C_2 \\ 0 & -\frac{\alpha}{\beta} \lambda^2 S_1 & C_0 & \beta S_1 \\ 0 & 0 & 0 & 1 \end{pmatrix}_i \{w\}_i \right. \tag{5'}$$

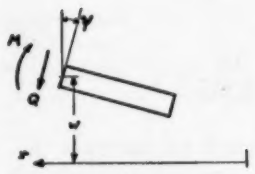


Fig. 1

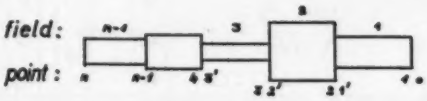


Fig. 2

Fig. 3

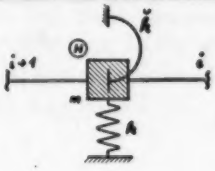


Fig. 4

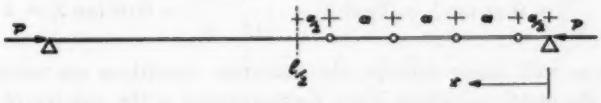


Fig. 6

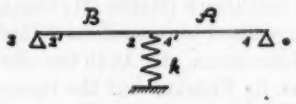


Fig. 5

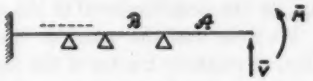


Fig. 7

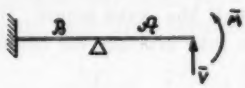


Fig. 8

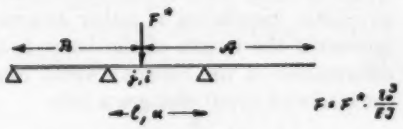


Fig. 9

we get for the problem of Fig. 2:

$$\{w\}_n = \mathfrak{A}_{n-1} \cdots \mathfrak{A}_1 \mathfrak{A}_1 \{w\}_0 = \mathfrak{P}\{w\}_0 \tag{6}$$

where $\mathfrak{A}$ stands for $\mathfrak{R}$ or $\mathfrak{K}$ and $\mathfrak{P}$ is the product matrix calculated according to the row × column rule. Including the actual boundary conditions, e.g., those of

Fig. 3, give

$$\begin{pmatrix} 0 \\ w_2 \\ 0 \\ w_4 \end{pmatrix}_n = \begin{pmatrix} \\ \mathfrak{P} \\ \\ \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ w_3 \\ w_4 \end{pmatrix}_0 .$$

Therefore, the equation for the eigenvalue is obtained from the condition that the determinant of the two equations

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} P_{13} & P_{14} \\ P_{23} & P_{24} \end{pmatrix} \begin{pmatrix} w_3 \\ w_4 \end{pmatrix}_0 \tag{7}$$

must vanish:

$$\begin{vmatrix} P_{13} & P_{14} \\ P_{23} & P_{24} \end{vmatrix} = 0 \tag{7'}$$

For $\mathfrak{A} = \mathfrak{R}$ and $\mathfrak{A} = \mathfrak{K}$ we find, for instance,

$$\begin{vmatrix} C_2 & S_2 \\ C_0 & S_1 \end{vmatrix} = 0 \text{ or } \tan\lambda = \operatorname{Tanh}\lambda \qquad \begin{vmatrix} C_2 & S_3 \\ C_0 & S_1 \end{vmatrix} = 0 \text{ or } \tan\lambda = \lambda \tag{7''}$$

For a beam with many sections, the transition conditions are automatically fulfilled by the matrix-notation. Thus, independently of the number of sections, of the presence of elastic supports, and also of the order of the eigenvalue, the characteristic equation is always a determinant of the second order. Their roots are determined by substitution of estimated eigenvalues in the region of interest, e.g., in the neighborhood of the angular speed of machinery (Holzer-Myklestad).

We note that the matrices $\mathfrak{R}$ and $\mathfrak{K}$ can be generalized to cover shear deformation, or rotatory inertia of the mass μ, or elastic foundation, etc. As to the vibrating beam, the first two generalizations were given by Fuhrke,[7] but the formulas are rather complicated for practical applications. Elastic foundation, on the other hand, offers no difficulty [Eq. (4) remains unchanged but $\omega^2\mu = (EI l^4/\lambda^4) + K$, where K denotes the elastic constant of the foundation]. For the column, an elastic foundation is rather awkward, but is fortunately without practical interest if the length of the beam is larger than half the wave length. Shear deformation of the column (which must be taken into account for beams of the sandwich type) changes $\mathfrak{K}$ into

$$\mathfrak{K}_s = \begin{pmatrix} 1 & (1+\lambda^2\zeta)S_1 & (1+\lambda^2\zeta)C_2 & (1+\lambda^2\zeta)(S_3 - \zeta S_1) \\ 0 & C_0 & S_1 & (1+\lambda^2\zeta)C_2 \\ 0 & -\lambda^2 S_1 & C_0 & (1+\lambda^2\zeta)S_1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

[7] *loc. cit.*

with

$$\zeta = \frac{EI}{GA_s l^2} \; (A_s = \text{shear area}) \text{ and } P = \frac{\lambda^2}{1 + \lambda^2 \zeta} \frac{EI}{l^2} \tag{4a}$$

3. Matrices for Points of Discontinuity: Supports, Single Masses, etc. If the structure contains concentrated supports, concentrated masses, pin connections, and the like, the procedure does not change, but special matrices are to be defined for such points. An elastic support k^* and a mass m have the same kind of effect on the column matrix $\{w\}$: the first three components remain continuous and V jumps by an amount proportional to w. In the same manner, a rotational spring $\check{k}$ and the moment of inertia Θ of a mass makes M jump proportionally to ψ. If Γ_V stands for $(k^* - m\omega^2)\,(l_0^3/EI_0)$, Γ_M for $(\check{k} - \Theta\omega^2)(l_0/EI_0)$ one gets

$$\{w\}_{i+1} = \quad \mathcal{S} \quad \{w\}_i \tag{8}$$

with

$$\mathcal{S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \Gamma_M & 1 & 0 \\ -\Gamma_V & 0 & 0 & 1 \end{pmatrix} \tag{8'}$$

At a pin connection M is continuous, but ψ is not.[8] If we introduce the influence number $\check{h}$ (reciprocal to an internal spring constant) and call $(EI_0/l_0)\check{h} = h$ the matrix is

$$\mathfrak{K} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & h & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{9}$$

We note that two other special matrices may be of interest: for the vibrating beam we define an "elastic matrix" $\mathcal{E}$ for those parts of the beam where the mass is comparatively small ($\mathcal{E} = \mathfrak{G}_{m=0}$), and for the column we introduce a matrix $\mathfrak{L}$ for the rigid parts between two elastic hinges ($\mathfrak{L} = \mathfrak{K}_{EI=\infty}$):

$$\mathcal{E} = \begin{pmatrix} 1 & \beta & \frac{1}{2}\frac{\beta^2}{\alpha} & \frac{1}{6}\frac{\beta^3}{\alpha} \\ 0 & 1 & \frac{\beta}{\alpha} & \frac{1}{2}\frac{\beta^2}{\alpha} \\ 0 & 0 & 1 & \beta \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathfrak{L} = \begin{pmatrix} 1 & \beta & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -p\beta & 1 & \beta \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{10}$$

[8] The case of discontinuity of w, being of less practical interest, is not discussed here.

The matrices $\mathfrak{S}$, $\mathfrak{K}$, $\mathfrak{E}$, $\mathfrak{L}$ with $p = Pl_0^2/EI_0$ are useful not only for beams where such special points occur but also for the approximate calculation of beams which in reality are continuous, for the transcendental matrices $\mathfrak{G}$ and $\mathfrak{K}$ are rather intricate as long as no special tables for the functions S_1, C_2 $\cdots$ etc. are available. And, as stated in 2, they get even less handy if shear deformation, etc. should be taken into account. In many cases it may therefore be advantageous to replace the continuity-matrices by discontinuity-matrices: for $\mathfrak{E}$ with shear deformation, only one term is changed

$$E_{14} \quad \text{from} \quad \tfrac{1}{6}\beta^3/\alpha \quad \text{to} \quad (\beta^3/\alpha)(\tfrac{1}{6} - \zeta) \quad \text{with} \quad \zeta = EI/GA_s l^2 \qquad (10')$$

A suitable number of matrices $\mathfrak{S}$ (with $\Gamma_M = 0$, $\Gamma_V = k = k^* l_0^3/EI_0$) may replace elastic foundations, and the matrices $\mathfrak{K}$ and $\mathfrak{L}$ permit the approximate treatment of columns as systems with a finite number of degrees of freedom. In such cases $\check{k}_i$ is defined by $(l/EI)_i$ and the pin connection is located in the middle of the piece of length l_i.

4. Systemization and Simplification of Calculations by use of "Δ-Matrices". The matrix edifice briefly described in the last sections has one weak spot. It cannot be used without additional considerations for the especially interesting limiting cases $\Gamma \to \infty$, $h \to \infty$, that is for systems with rigid supports or ideal hinges. To illustrate this problem, consider the example of Fig. 5. In this instance, the equation for the eigenvalue is

$$\begin{vmatrix} B_{1\nu}A_{\nu 2} - kB_{14}A_{12} & B_{1\mu}A_{\mu 4} - kB_{14}A_{14} \\ B_{3\nu}A_{\nu 2} - kB_{34}A_{12} & B_{3\mu}A_{\mu 4} - kB_{34}A_{14} \end{vmatrix} = c_0 k^2 + c_1 k + c_2 \qquad (11)$$

(Summation over the repeated indices.) If we carry out the calculations of the coefficients c_i, it appears that $c_0 = 0$, and that half of the sixteen c_1-products and eight of the thirty two c_2-products vanish. In numerical computations, therefore, the highest power of k is multiplied by a small number (actually 0), and if k is large this leads to considerable numerical errors; furthermore, the especially simple case of rigid supports ($k = \infty$) is numerically unsolvable by direct calculation in this way. One is obliged to transport k, h, etc. as parameters throughout the whole calculation, an obviously very inconvenient procedure.

The difficulties thus arising can be overcome by a device that makes the calculation of the final determinant more straightforward. We arrange the thirty six (2×2) subdeterminants of a (4×4) matrix again into a matrix scheme[9], in other words adjoin to the 4×4 matrices $\mathfrak{A}$, $\mathfrak{B}$ the 6×6 "Δ_2-matrices", $\mathfrak{A}^{(2)}$, $\mathfrak{B}^{(2)}$. It then can be shown[10] that the well-known scalar equation

$$|\mathfrak{B}\mathfrak{A}| = |\mathfrak{B}|\,|\mathfrak{A}|$$

for any two $m \times m$-matrices can be generalized to the matrix equation

[9] Fuhrke, *loc. cit.* p. 335.

[10] A proof is given in the thesis of Fuhrke [footnote 1] who first introduced this idea into calculations of that kind. The mathematical lemma itself, as we found out recently is well known, s.e.g. Hasse: Aufgabensammlung zur Höh Algebra, Göschen Nr 1082, p. 57.

$$(\mathfrak{B}\mathfrak{A})^{(2)} = \mathfrak{B}^{(2)}\mathfrak{A}^{(2)} \tag{12}$$

valid for any subdeterminant "Δ" of order $\leqq m$. This means that the process of passing to the determinant can be anticipated and it turns out that by this procedure the terms finally cancelling out do not appear at all. Thus, fewer terms need to be calculated, and the differences of large numbers cannot occur.

The process of adjoining the Δ-matrix to the original matrix is to a certain extent arbitrary. We adopt the following simple "dictionary"-rule for coordinating the combinations of rows and columns of $\mathfrak{A}$ to the rows and columns of $\mathfrak{A}^{(2)}$: The combinations

$$\begin{array}{cccccc} 12 & 13 & 14 & 23 & 24 & 34 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{array} \tag{12'}$$

are called

i.e.,

$$A_{11}^{(2)} = \begin{vmatrix} A_{11} A_{12} \\ A_{21} A_{22} \end{vmatrix}, \quad A_{12}^{(2)} = \begin{vmatrix} A_{11} A_{13} \\ A_{21} A_{23} \end{vmatrix} \cdots A_{35}^{(2)} = \begin{vmatrix} A_{12} A_{14} \\ A_{42} A_{44} \end{vmatrix} \quad \text{etc.}$$

With the theorem (12) the problem Fig. 5 can now be restated as

$$0 = (\mathfrak{B}\mathfrak{S}\mathfrak{A})_{26}^{(2)} = (\mathfrak{B}^{(2)}\mathfrak{S}^{(2)}\mathfrak{A}^{(2)})_{26}$$

$$= \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ B_{21}^{(2)} & B_{22}^{(2)} & B_{23}^{(2)} & B_{24}^{(2)} & B_{25}^{(2)} & B_{26}^{(2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ k & 0 & 0 & 0 & 1 & 0 \\ 0 & k & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & A_{16}^{(2)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{26}^{(2)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{36}^{(2)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{46}^{(2)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{56}^{(2)} & \cdot \\ \cdot & \cdot & \cdot & \cdot & A_{66}^{(2)} & \cdot \end{pmatrix} \tag{13}$$

or

$$B_{2\rho}^{(2)} A_{\rho 5}^{(2)} + k(B_{25}^{(2)} A_{15}^{(2)} + B_{26}^{(2)} A_{25}^{(2)}) = 0$$

(Only the explicitly written matrix elements need be calculated.) Here the case $k = \infty$ offers no difficulty: for 3 rigid supports we have

$$B_{2(\rho+1)} A_{\rho 5} = 0 \qquad \rho = 1, 2 \tag{14}$$

Accordingly we get

$$D_{2(4+\tau)} C_{\tau(4+\sigma)} B_{\sigma(4+\rho)} A_{\rho 5} = 0 \qquad \tau, \sigma, \rho = 1, 2$$

for, e.g., five supports. This formula can be derived from a formula analagous to (13) with $k = \infty$, and dividing by k throughout, i.e., writing

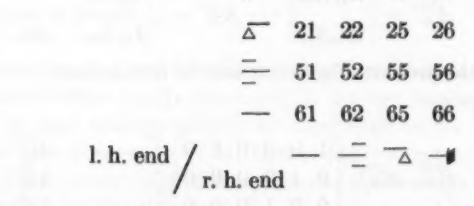
$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} = \mathfrak{S}_r^{(2)}$$

instead of $\mathfrak{S}^{(2)}$ between the $\mathfrak{A}^{(2)}$, $\mathfrak{B}^{(2)}$ $\cdots$ ("r" = rigid).

The difficulty mentioned in discussing Eq. (11) is an example of a very general statement: If one calculates a product $\cdots \mathfrak{D}\mathfrak{C}\mathfrak{B}\mathfrak{A}$, every element $c_{\mu\nu}$, $l_{\lambda\lambda}$ of the

matrices $\cdots$ $\mathfrak{C}$, $\mathfrak{G}$ appears in every element of the product. These elements are therefore squared when we build the final determinant. But from (12), most generally, it is seen that the final determinant is a linear function of all the constituent elements. This means that, since the factors of $l_{\lambda\lambda}^2 \cdots$ never disappear exactly in numerical computations, any large value $l_{\lambda\lambda}$ compromises the accuracy of numerical work.

If the boundary conditions of a problem are to be specified in advance, the final Δ-matrix can be regarded as a scheme where each element provides the solution for a special combination of boundary conditions. Obviously not all combinations are mechanically possible; $M = 0$ contradicts $\psi = 0$, and so do $V = 0$ and $w = 0$. The remaining elements are illustrated below, the boundary conditions being symbolized in the order: end- 1) clamped, 2) hinged, 3) clamped but not pinned, 4) free.



(15)

The other elements of the Δ_2-matrix are necessary for the computation of those indicated if products of several Δ_2-matrices have to be considered.

5. List of Δ_2-Matrices. To apply Δ_2-matrices for practical purposes one needs a list of such matrices. Such a list is given here:

$$\mathfrak{R}^{(2)} = \begin{pmatrix} \frac{1}{2}(C_0^* + 1) & S_1^* & C_2^* & C_2^* & S_3^* & \frac{1}{2}(1 - C_0^*)\frac{1}{\lambda^4} \\ -\lambda^4 S_3^* & C_0^* & S_1^* & S_1^* & 2C_2^* & S_3^* \\ -\lambda^4 C_2^* & -\lambda^4 S_3^* & \frac{1}{2}(C_0^* + 1) & \frac{1}{2}(C_0^* - 1) & S_1^* & C_2^* \\ -\lambda^4 C_2^* & -\lambda^4 S_3^* & \frac{1}{2}(C_0^* - 1) & \frac{1}{2}(C_0^* + 1) & S_1^* & C_2^* \\ -\lambda^4 S_1^* & -2\lambda^4 C_2^* & -\lambda^4 S_3^* & -\lambda^4 S_3^* & C_0^* & S_1^* \\ \frac{1}{2}\lambda^4(1 - C_0^*) & -\lambda^4 S_1^* & -\lambda^4 C_2^* & -\lambda^4 C_2^* & -\lambda^4 S_3^* & \frac{1}{2}(C_0^* + 1) \end{pmatrix} \tag{16}$$

$$\mathfrak{K}^{(2)} = \begin{pmatrix} C_0 & S_1 & C_2 & C_2 & S_3^* & C_4^* \\ -\lambda^2 S_1 & C_0 & S_1 & S_1 & S_1 & S_3^* \\ 0 & 0 & 1 & 0 & S_1 & C_2 \\ 0 & 0 & 0 & 1 & S_1 & C_2 \\ 0 & 0 & 0 & 0 & C_0 & S_1 \\ 0 & 0 & 0 & 0 & -\lambda^2 S_1 & C_0 \end{pmatrix}$$

with—in $\mathfrak{R}^{(2)}$—

$$C_0^* = \quad \text{Cosh } \lambda \cos \lambda \quad = 1 - \frac{1}{4!}(\lambda\sqrt{2})^4 + \cdots$$

$$S_1^* = \frac{1}{2\lambda}(\text{Cosh } \lambda \sin \lambda + \text{Sinh } \lambda \cos \lambda) = \frac{1}{1!} - \frac{1}{5!}(\lambda\sqrt{2})^4 + \cdots$$

$$C_2^* = \frac{1}{2\lambda^2} \quad \text{Sinh } \lambda \sin \lambda \quad = \frac{1}{2!} - \frac{1}{6!}(\lambda\sqrt{2})^4 + \cdots$$

$$S_3^* = \frac{2}{4\lambda^3}(\text{Cosh } \lambda \sin \lambda - \text{Sinh } \lambda \cos \lambda) = 2\left(\frac{1}{3!} - \frac{1}{7!}(\lambda\sqrt{2})^4 + \cdots\right) \qquad (16')$$

and—in $\mathfrak{K}^{(2)}$—C_0, S_1, C_2, S_3 as before and

$$S_2^* = \frac{1}{\lambda^2}(\sin \lambda - \lambda \cos \lambda) = \frac{1}{\lambda^2}(S_1 - C_0)$$

$$C_4^* = \frac{1}{\lambda^4}(2(1 - \cos \lambda) - \lambda \sin \lambda) = \frac{1}{\lambda^2}(2C_2 - S_1)$$

If, for a beam of many sections, the quantities α, β from Eq. (5) ought to be introduced (see Eq. (5′)) the elements of the matrices (16) must be multiplied *element by element* by the scheme

$$\begin{pmatrix} 1 & \frac{\beta}{\alpha} & \frac{\beta^2}{\alpha} & \frac{\beta^2}{\alpha} & \frac{\beta^3}{\alpha} & \frac{\beta^4}{\alpha^2} \\ \frac{\alpha}{\beta} & 1 & \beta & \beta & \beta^2 & \frac{\beta^3}{\alpha} \\ \frac{\alpha}{\beta^2} & \frac{1}{\beta} & 1 & 1 & \beta & \frac{\beta^2}{\alpha} \\ \frac{\alpha}{\beta^2} & \frac{1}{\beta} & 1 & 1 & \beta & \frac{\beta^2}{\alpha} \\ \frac{\alpha}{\beta^3} & \frac{1}{\beta^2} & \frac{1}{\beta} & \frac{1}{\beta} & 1 & \frac{\beta}{\alpha} \\ \frac{\alpha^2}{\beta^4} & \frac{\alpha}{\beta^3} & \frac{\alpha}{\beta^2} & \frac{\alpha}{\beta^2} & \frac{\alpha}{\beta} & 1 \end{pmatrix} \qquad (16'')$$

The functions for the vibrating beam are essentially the functions of Hohenemser and Prager,[11] slightly modified in the light of their new interpretation. The symbols are altered according to the fact that the four functions can be interpreted as Rayleigh-functions of the argument $(1 + i)$, $(\text{Cosh}(1 + i)\lambda + \cos(1 + i)\lambda = 2 \text{ Cosh } \lambda \cos \lambda$ etc.). The well-converging series may be helpful for numerical computations.

The matrices $\mathfrak{R}^{(2)}$ and $\mathfrak{K}^{(2)}$ are cross-symmetrical, and from an interpretation of the matrices as schemes of influence-numbers it can be seen that the equalities

[11] Dynamik der Stabwerke, Berlin 1933, p. 127.

$A_{n3} = A_{n4}(n = 1, 2, 5, 6)$ $A_{3m} = A_{4m}(m = 1, 2, 5, 6)$ are the consequences of Maxwell's reciprocity theorem (see 7).

The Δ_2-matrix for supports, or for concentrated masses is (see Eq. 8′):

$$\mathfrak{S}^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \Gamma_M & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \Gamma_V & 0 & 0 & 0 & 1 & 0 \\ \Gamma_V\Gamma_M & \Gamma_V & 0 & 0 & \Gamma_M & 1 \end{pmatrix} \tag{17}$$

Furthermore

$$\mathfrak{K}^{(2)} = \begin{pmatrix} 1 & h & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & h \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{18}$$

and

$$\mathfrak{E}^{(2)} = \left(\begin{array}{cccccc} 1 & \frac{\beta}{\alpha} & \frac{\beta^2}{2\alpha} & \frac{\beta^3}{2\alpha} & \frac{\beta^3}{\alpha}\left(\frac{1}{3}+\zeta\right) & \frac{\beta^4}{\alpha}\left(\frac{1}{12}+\zeta\right) \\ 0 & 1 & \beta & \beta & \beta^2 & \frac{\beta^3}{\alpha}\left(\frac{1}{3}+\zeta\right) \\ 0 & 0 & 1 & 0 & \beta & \frac{\beta^2}{2\alpha} \\ 0 & 0 & 0 & 1 & \beta & \frac{\beta^2}{2\alpha} \\ 0 & 0 & 0 & 0 & 1 & \frac{\beta}{\alpha} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \tag{19}$$

$$\mathfrak{L}^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -p\beta & 1 & \beta & \beta & \beta^2 & 0 \\ 0 & 0 & 1 & 0 & \beta & 0 \\ 0 & 0 & 0 & 1 & \beta & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -p\beta & 1 \end{pmatrix}$$

In the case of the vibrating beam with no distributed mass [$\mathfrak{R}^{(2)} \to \mathfrak{E}^{(2)}$], and also for the column considered as a chain of rigid bars,[12] it may be useful to give the matrices which combine $\mathfrak{S}^{(2)}$ and $\mathfrak{E}^{(2)}$, or $\mathfrak{L}^{(2)}$ and $\mathfrak{K}^{(2)}$ (written for convenience as

[12] Some information about this idea can be found in Biezeno-Grammel, "Technische Dynamik", 1st ed., pp. 172, 509.

the sums of matrices):

$$\mathcal{S}^{(2)}\mathcal{E}^{(2)} = \mathcal{E}^{(2)} + \Gamma_V \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ [E_1] \\ [E_2] \end{pmatrix} + \Gamma_M \begin{pmatrix} 0 \\ [E_1] \\ 0 \\ 0 \\ 0 \\ [E_3] \end{pmatrix} + \Gamma_V\Gamma_M \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ [E_1] \end{pmatrix} \quad \mathfrak{L}^{(2)}\mathfrak{K}^{(2)} = \begin{pmatrix} 1 & h & 0 & 0 & 0 & 0 \\ 0 & 1 & \beta & \beta & \beta^2 & \beta^2 h \\ 0 & 0 & 1 & 0 & \beta & \beta h \\ 0 & 0 & 0 & 1 & \beta & \beta h \\ 0 & 0 & 0 & 0 & 1 & h \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} - p\beta \begin{pmatrix} 0 \\ [H_1] \\ 0 \\ 0 \\ 0 \\ [H_3] \end{pmatrix} \tag{20}$$

$[E_1] \cdots [H_3]$ etc., indicate the first, $\cdots$ fifth row of the $\mathcal{E}^{(2)}$-, $\mathfrak{K}^{(2)}$-matrix. Note the simplifications that come with $\Gamma_V = \infty$ etc. and which, as in the example (14), can most generally be introduced before the start of numerical computations.

In the especially interesting case of $k^* = \check{k} = 0$ and of masses with negligible moments of inertia ($\Gamma_V = -m\omega^2 l_0^3/EI_0 \equiv -q$, $\mathcal{S}^{(2)} \to \mathfrak{M}^{(2)}$) the product $\mathcal{S}^{(2)}\,\mathcal{E}^{(2)}$ simplifies so that we have two formulae analogous to each other:

$$\mathfrak{M}^{(2)}\mathcal{E}^{(2)} = \mathcal{E}^{(2)} - q \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ [E_1] \\ [E_2] \end{pmatrix} \qquad \mathfrak{L}^{(2)}\mathfrak{K}^{(2)} = \mathcal{E}_r^{(2)} - p\beta \begin{pmatrix} 0 \\ [H_1] \\ 0 \\ 0 \\ 0 \\ [H_3] \end{pmatrix} \tag{20'}$$

$\mathcal{E}_r^{(2)}$, explicitly given in (20), is a degenerated matrix $\mathcal{E}^{(2)}$, all bending elasticity being concentrated at the right-hand end; $\mathcal{E}_r^{(2)}$, naturally, is not cross-symmetrical.

The formulae (20′) show that for simple chains of n concentrated masses or n rigid columns the equation for the eigenvalue can immediately be written down as an equation of the n-th order, the coefficients being products of matrices which do not contain the eigenvalue and therefore (contrary to the case of the trancendental matrices) need be caluclated only once.

6. An Example. To illustrate the procedure based on (20) we calculate the Euler-load for a hinged-hinged column which we replace by a chain according to Fig. 6. We take equally spaced hinges, $a = l/n$, $h = a/EI$, and make the distance between the first hinge and $x = 0$ equal to $a/2$.[13] $l_0 = a$ and $EI_0 = EI$ are length and stiffness of references. For reasons of symmetry we need regard one one-half of the beam, say between 0 and $l/2$. The boundary conditions for the product-Δ_2-matrix are then "55", because on the left-hand side w_2 and w_4 are zero, and on the right-hand side w_2 and w_4 are unknown.[14] The calculation runs as follows:

[13] For our purpose there is no advantage in using different values of h as discussed in Biezeno-Grammel, pp. 172, 512.

[14] With this formulation the hinge $x = 0$ is regarded as "the" boundary (condition: $M = 0$), and between this point and the first hinge with h we have the matrix $\mathfrak{L}(a)$ (19). But the same result of course comes out, if one regards the hinge $x = 0$ as a part of the beam (take now (20′) with $h = \infty$ instead of (19)), and formulate as boundary conditions the fact that $M = V = 0$ immediately right of $x = 0$. Then the boundary conditions of our problem would be "56". This second interpretation is very useful in many cases, e.g., for elastically hinged beams, but here it is more convenient to stay with the first one.

We multiply the matrices from the left to the right, starting with the fifth row of the Δ_2-matrix for the piece in the middle of the length $a/2$. From (20) it is seen that this row is simply $[0, 0, 0, 0, 1, h]$; $p = Pa^2/EI$ does not appear because this piece lies in the direction of P. The product of this row with the next matrix (20) $[\alpha = 1, \beta = 1]$ gives a new fifth row $[0, 0, 0, 0, 1, 2h] - ph[0, 0, 0, 0, 1, h]$ the next step leads to another fifth row.

$$[0, 0, 0, 0, 1, 3h] - ph[0, 0, 0, 0, 3, 4h] + (ph)^2[0, 0, 0, 0, 1, h]$$

etc. The last step would be multiplication by the fifth column of $\mathfrak{L}^{(3)}(19)$ (with $\beta = \frac{1}{2}$), i.e., multiplication of the fifth element by 1, of the sixth element by $-p/2$. With z standing for ph we thus get the following equations for n (number of hinges) $= 2, 4, 6, 8, 10$:

$$\begin{aligned}
n &= 2 \quad & 1 - \tfrac{1}{2}z &= 0 \\
n &= 4 \quad & 1 - 2z + \tfrac{1}{2}z^2 &= 0 \\
n &= 6 \quad & 1 - \tfrac{5}{2}z + 3z^2 + \tfrac{1}{2}z^3 &= 0 \\
n &= 8 \quad & 1 - 8z + 10z^2 - 4z^3 + \tfrac{1}{2}z^4 &= 0 \\
n &= 10 \quad & 1 - \tfrac{25}{2}z + 25z^2 - \tfrac{35}{3}z^3 + 5z^4 - \tfrac{1}{2}z^5 &= 0
\end{aligned} \qquad (21)$$

The reasons for writing down these equations is to show that their solution offers no difficulty, because from 3 terms a very good first approximation can always be found. The numerical results for $Pl^2/EI = n^2z$ are:

$n = 2$		4	6	8	10
first approx.	8	9.38	9.77	9.92	10.0 converging
final value	8	9.38	9.65	9.77	9.80 to $\pi^2 = 9.87$

If one wishes to avoid trancendental equations, and if a rough estimation is sufficient, this procedure, taking only a few hinges, may be satisfactory, especially if the eigen-shape does not contain inflection points.

7. Inhomogeneous Equations. The Δ-matrix notation is very convenient, not only for determining eigenvalues, but also for the calculation of deflection under lateral loads (forced vibrations,[15] column-beam problems). The example Fig. 7 where $\mathcal{P}$ stands for the product of the matrices $\cdots$ $\mathfrak{B}\mathfrak{A}$ may illustrate this fact. At the right-hand end a moment $\bar{M} = EI_0/l_0^2\,\overline{w}_3$ and a vertical force $\bar{V} = (EI_0/l_0^3)\,\overline{w}_4$ may act. The left-hand boundary conditions may be specified by any of the possibilities

m, n	12	13	24	34
r	1	2	5	6

then from

$$\left.\begin{aligned} 0 &= P_{m1}w_0 + P_{m2}l_0\psi_0 + P_{m3}\overline{w}_3 + P_{m4}\overline{w}_4 \\ 0 &= -P_{n1}w_0 + P_{n2}l_0\psi_0 + P_{n3}\overline{w}_3 + P_{n4}\overline{w}_4 \end{aligned}\right\}$$

[15] Fuhrke, *loc. cit.*

there follows

$$\left.\begin{aligned} w_0 &= \frac{P^{(2)}_{r4}}{P^{(2)}_{r1}}\,\overline{w}_3 + \frac{P^{(2)}_{r6}}{P^{(2)}_{r1}}\,\overline{w}_4 \\ -l_0\psi_0 &= \frac{P^{(2)}_{r2}}{P^{(2)}_{r1}}\,\overline{w}_3 + \frac{P^{(2)}_{r3}}{P^{(2)}_{r1}}\,\overline{w}_4 \end{aligned}\right\} \tag{22}$$

The coefficients of $\overline{w}_3$ and $\overline{w}_4$ are influence coefficients for displacement and rotation at the free end. From Maxwell's theorem it follows that

$$P^{(2)}_{r4} = P^{(2)}_{r3} \tag{23}$$

as announced above[4]. In the case of Fig. 8, e.g., the calculation of the elements $P^{(2)}_{pq}$ gives

$$\mathfrak{P}^{(2)} = \mathfrak{B}^{(2)}\mathfrak{S}^{(2)}_{r1g}\,\mathfrak{A}^{(2)} = \mathfrak{B}^{(2)} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ [A_1] \\ [A_2] \end{pmatrix}$$

or

$$\left.\begin{aligned} P^{(2)}_{11} &= B^{(2)}_{15}A^{(2)}_{11} + B^{(2)}_{16}A^{(2)}_{21}, \qquad P^{(2)}_{14} = B^{(2)}_{15}A^{(2)}_{14} + B^{(2)}_{16}A^{(2)}_{24}, \\ P^{(2)}_{15} &= B^{(2)}_{15}A^{(2)}_{15} + B^{(2)}_{16}A^{(2)}_{25} \text{ etc.} \end{aligned}\right\} \tag{24}$$

a procedure which may prove to be useful even for mere statics ($\mathfrak{A}^{(2)}$, $\mathfrak{B}^{(2)} \to \mathfrak{E}^{(2)}$).

In the more general case of a force acting at an intermediate point, i, $i+1$ (e.g. Fig. 9) we get (again with $\mathfrak{P} = \mathfrak{B}\mathfrak{A}$):

$$\left.\begin{aligned} -w_{(0)} &= \frac{F}{P^{(2)}_{12}}\,(P_{12}B_{34} - P_{32}B_{14}) = \frac{F}{P^{(2)}_{12}}\,(B^{(2)}_{23}A_{12} + B^{(2)}_{26}A_{22} + B^{(2)}_{28}A_{32}) \\ l_0\psi_{(0)} &= \frac{F}{P^{(2)}_{12}}\,(P_{11}B_{34} - P_{31}B_{14}) = \frac{F}{P^{(2)}_{12}}\,(B^{(2)}_{23}A_{11} + B^{(2)}_{26}A_{21} + B^{(2)}_{28}A_{31}) \end{aligned}\right\} \tag{25}$$

$w_{(0)}, l_0\psi_{(0)}$ being known the calculations of any value $w(x) \cdots V(x)$ offers no difficulty. Note that rigid supports are included in (25), if we divide by k throughout and put $k = \infty$.

8. Use of Δ-Matrices in Computational Work. The use of Δ-matrices is obviously profitable if the equation for the eigenvalue can be written in general terms; for an elimination in advance of unnecessary terms is clearly very advantageous. The advantage of the matrices is less pronounced if, for rather complex problems, only numerical procedures can be used. It can easily be seen that, e.g., the number of multiplications necessary for the computation of the final determinant is as follows:

Δ_2-matrices	matrices
$6^2(n-2)+6 = 36(n-2)+6$	$4^2(n-2)2+4\cdot 4 = 32(n-2)+16$

(26)

and it is obvious how simple matrix multiplication now gets. Two fields lead to

$$\cos \lambda_2 \beta_1 \frac{\sin \lambda_1}{\lambda_1} + \cos \lambda_1 \beta_2 \frac{\sin \lambda_2}{\lambda_2} = 0$$

or

$$t_1 + t_2 = 0 \quad \text{with} \quad t_i = \beta_i \frac{\tan \lambda_i}{\lambda_i}.$$

In the case of 3 fields we find

$$\left(\cos \lambda_2 \cos \lambda_2 - \frac{\beta_3 \lambda_2}{\beta_2 \lambda_3} \sin \lambda_3 \sin \lambda_2\right) \beta_1 \frac{\sin \lambda_1}{\lambda_1}$$

$$+ \left(\cos \lambda_3 \beta_2 \frac{\sin \lambda_2}{\lambda_2} + \beta_3 \frac{\sin \lambda_3}{\lambda_3} \cos \lambda_2\right) \cos \lambda_1 \quad \text{or} \quad t_1 + t_2 + t_3 = \left(\frac{\lambda_2}{\beta_2}\right)^2 t_1 t_2 t_3$$

In the case of 4 fields the result can be written as

$$t_1 + t_2 + t_3 + t_4 = \left(\frac{\lambda_2}{\beta_2}\right)^2 t_1 t_2 (t_3 + t_4) + \left(\frac{\lambda_3}{\beta_3}\right)^2 t_3 t_4 (t_1 + t_2) \quad \text{etc.}$$

An analogous simplification is feasible for the cantilever-column (1st Euler case): V being zero, and w without influence as long as only one point is fixed, the matrix (4) reduces to

$$\begin{pmatrix} l_0 \psi \\ \dfrac{l_0^2}{EI} M \end{pmatrix}_1 = \begin{pmatrix} \cos \lambda & \dfrac{\beta}{\alpha} \dfrac{\sin \lambda}{\lambda} \\ -\dfrac{\alpha}{\beta} \lambda \sin \lambda & \cos \lambda \end{pmatrix} \begin{pmatrix} l_0 \psi \\ \dfrac{l_0^2}{EI_0} M \end{pmatrix}_0.$$

TECHNISCHE HOCHSCHULE
DARMSTADT, GERMANY

so that for $n > 4$ the Δ_2-matrices require slightly more computational work than the direct way. But this difference is over-compensated by the fact that the "Δ"-way involves no dangerous differences of large numbers. And for very large values of n a combined procedure can be used: at parts of the beam, where no highly concentrated masses or stiffnesses act, the ordinary matrix multiplication is used. But for the connection of these parts over stiff supports, large masses, etc., the Δ_2-matrices should be used to avoid the inaccurate differences.

The mixed procedure which in the case of a 4×4 matrix problem is remunerative only for very large values of n, becomes a necessity for problems of higher order. If, for example, bending and torsion are coupled, the governing differential equation is of the sixth order, and the resulting matrix therefore of 6×6 order. As three boundary conditions ought to be fulfilled at both ends of the beam, we are interested in determinants of the third order. But a 6×6 matrix has 20×20 subdeterminants of the third order ($\Delta^{(3)}$-matrix; generally: an $m \times m$ matrix has $\binom{m}{k}^2$ subdeterminants of the k^{th} order), and therefore the number of multiplications increases very rapidly, e.g.,

$$20^2(n-2)+20 \quad \text{vs} \quad 3\cdot 6^2(n-2)+6\cdot 9 \tag{27}$$

The procedure should therefore be as follows: Calculate $\mathfrak{A}^{(3)}$, $\mathfrak{B}^{(3)}$, etc. for every "field" in the normal way with, of course, only those elements which are needed for the actual boundary conditions; then multiply these matrices, intercalating the $\mathfrak{S}^{(3)}$-matrix for the supports. If there are rigid supports, the $\mathfrak{S}^{(3)}$-matrix is highly degenerate so that fewer multiplications occur as Eq. (27) indicates.[16]

9. Conclusion. This paper shows that matrices are very useful for the exact calculation of eigenvalues of vibrating and buckling beams with varying cross-section and intermediate supports. It further shows how calculations can be made shorter and safer by introducing auxiliary matrices, called Δ-matrices. A listing of all matrices eventually of interest for such calculations is also given.

10. Appendix. Matrix Solution for an n-Section Beam. In the special case of the hinged-hinged beam (2nd Euler case), the calculation of an n-section column,[17] a typical matrix problem, can be simplified by starting with the equation

$$EIw'' + Pw = 0$$

instead of (1). The matrix formula analogous to Eq. (4) is

$$\begin{pmatrix} w \\ l_0\psi \end{pmatrix}_1 = \begin{pmatrix} \cos\lambda & \beta\dfrac{\sin\lambda}{\lambda} \\ \dfrac{\lambda\sin\lambda}{\beta} & \cos\lambda \end{pmatrix} \begin{pmatrix} w \\ l_0\psi \end{pmatrix}_0$$

[16] An example of frame vibrations evaluated in the above-mentioned thesis of Fuhrke (Ing. Arch. 1956) shows that the Δ_3-matrices can be of practical value.

[17] Thomson, Jour. Appl. Mech. 1950, p. 132 avoids the much handier matrix notation.

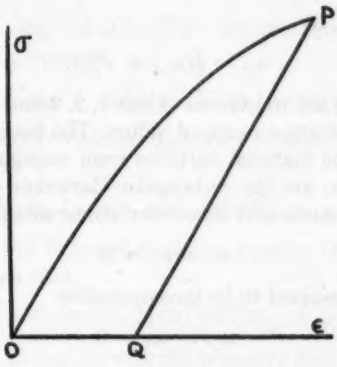


FIG. 1

of σ. If after σ has reached a certain value, unloading takes place, Eq. (2.1) is no longer valid and a linear relation is found to exist between σ and ϵ; the ratio of the differentials of stress and strain is constant. For a complete unloading the specimen displays a permanent extension given by OQ in Fig. 1. Thus for simple tension (or compression) the governing stress-strain relations for the stress theory of plastic flow are, for loading and unloading respectively,

$$d\sigma = 3G_0\, d\epsilon - (3a_1\epsilon + 5a_2\epsilon^3 + \cdots)\epsilon\, d\epsilon \tag{2.2}$$

$$d\sigma = 3G_0\, d\epsilon \tag{2.3}$$

where Eq. (2.2) is the differential form of Eq. (2.1).

A possible method of establishing an analytic loading criterion in terms of strain is as follows: under tension $d\epsilon > 0$, under compression $d\epsilon < 0$, under either $\epsilon\, d\epsilon = d(\epsilon^2/2) > 0$. A somewhat more convenient form can be obtained by changing Eq. (2.2) to

$$d\sigma = 3G_0\, d\epsilon - (b_1\epsilon + b_2\epsilon^3 + \cdots)(\epsilon + \epsilon^3 + \cdots)\, d\epsilon \tag{2.4}$$

where $b_1 = 3a_1$, $b_2 = 5a_2 - 3a_1$, etc. Here

$$(\epsilon + \epsilon^3 + \cdots)\, d\epsilon = d(\epsilon^2/2 + \epsilon^4/4 + \cdots) > 0$$

can serve as the loading criterion. For unloading, when $d(\epsilon^2/2 + \epsilon^4/4 + \cdots) < 0$, the stress-strain relation (2.3) is to be used. It should be noted that for the case when $d(\epsilon^2/2 + \epsilon^4/4 + \cdots) = 0$, which has no physical meaning in this test, the two stress-strain relations (2.3) and (2.4) are identical.

Generalizations for combined states of stress and strain involve the introduction of the stress and strain deviations, s_{ij} and e_{ij} respectively, defined as

$$s_{ij} = \sigma_{ij} - \tfrac{1}{3}\delta_{ij}\sigma_{kk} \tag{2.5}$$

where σ_{ij} is the stress tensor, δ_{ij} the Kronecker delta, and

$$e_{ij} = \epsilon_{ij} - \tfrac{1}{3}\delta_{ij}\epsilon_{kk} \tag{2.6}$$

where ϵ_{ij} is the strain tensor given, for moderate deformations to which this

STRESS THEORY OF PLASTIC FLOW

D. TRIFAN

1. Introduction. This paper is concerned with a mathematical theory of plastic flow for incompressible, isotropic, strain-hardening materials exhibiting a gradual transition from the elastic to the plastic state. A flow theory for materials of this type was introduced by Prager [1][1], and later generalized by Handelman, Lin, and Prager [2], where the differential of strain is dependent on the state of stress and the differential of stress. One may refer to this theory as a strain theory of plastic flow. For a certain set of problems this type of stress-strain relation is not in a suitable form, and for this reason another flow theory, referred to as a stress theory of plastic flow, was introduced [3] where the differential of stress is dependent on the state of strain and the differential of strain. Prandtl [4] and Reuss [5] introduced still another type of flow theory where the differential of stress is dependent on the state of stress and the differential of strain, however, this theory is not valid for materials under consideration here, but for perfectly plastic materials. It is the purpose of this paper to generalize the stress theory of plastic flow as given in [3] and [6].

The general stress-strain relations are developed in Sec. 2 subject to the conditions of incompressibility, isotropy, and gradual transition from the elastic to the plastic state. In addition, these relations are also subject to the continuity condition which requires identical differentials of stress at the boundary between the domains of loading and unloading. In Sec. 3 the inverted form of the stress-strain relations is obtained while in Sec. 4 the conditions on the arbitrary functions of the stress-strain law are introduced. Two absolute minimum principles are given in Sec. 5 while the last section emphasizes some limitations of the theory.

2. General stress-strain relations. In the discussion to follow only states of stress and strain which can be obtained by one loading followed by a partial or complete unloading will be considered. For a simple tension test where σ denotes the tensile stress and ϵ the corresponding unit extension, this can be illustrated by the following stress-strain diagram where the functional relation between σ and ϵ for loading is given by the curve OP and for unloading by the line PQ.

A stress theory of plastic flow for this simple state of stress and strain, the results of which can be used as a model for a general stress-strain relation, can be developed by assuming the finite relation between σ and ϵ for loading as

$$\sigma = 3G_0\epsilon - (a_1\epsilon^3 + a_2\epsilon^5 + \cdots), \tag{2.1}$$

where G_0 is the elastic shear modulus, and the remaining coefficients are positive constants which are also dependent on the material. The form of Eq. (2.1) assumes that under simple compression the stress-strain diagram will be symmetrical with respect to the origin, i.e., a change of sign of ϵ merely changes the sign

[1] Numbers in brackets refer to the bibliography at the end of the paper.

discussion is restricted, by

$$\epsilon_{ij} = \tfrac{1}{2}(u_{i,j} + u_{j,i}). \tag{2.7}$$

The free indices i and j are to take on values 1, 2, 3 and the repeated indices are to be summed over the same range of values. The function u_i denotes the displacement vector of the material particles from standard state, and $u_{i,j}$ represents $\partial u_i/\partial x_j$, where x_j are the rectangular Cartesian coordinates. As a result of the definitions, the stress and strain deviations satisfy

$$s_{ii} = e_{ii} = 0. \tag{2.8}$$

Since the material is assumed to be incompressible

$$u_{i,i} = \epsilon_{ii} = 0. \tag{2.9}$$

It follows firstly that $\epsilon_{ij} = e_{ij}$, and secondly that stress-strain relations are dependent on stress deviations alone[2].

An extension of the stress-strain relation for unloading given by (2.3) can be obtained by assuming that the differential form of the generalized Hooke's law for incompressible solids is applicable, i.e. an elastic behavior. Expressed in terms of stress and strain deviations, the stress-strain relation for unloading, for combined states of stress and strain, becomes

$$ds_{ij} = 2G_0\, de_{ij}. \tag{2.10}$$

For an extension of the loading stress-strain relation given by (2.4) the following assumptions will be made: 1) the differential of stress deviation is a linear function in the differential of strain deviation as is indicated by (2.4); 2) the principal axes of stress and strain coincide—due to the isotropy condition; 3) the factor $(b_1\epsilon + b_2\epsilon^3 + \cdots)$ is replaced by a series of odd powered strain deviations, say K_{ij}; 4) the differential factor $d(\epsilon^2/2 + \epsilon^4/4 + \cdots)$ is replaced by a differential of a series of strain deviation invariants of even order, say dg; and 5) the stress-strain relation reduces to (2.10) for neutral change of strain. The most general form of (2.4) consistent with these conditions is

$$ds_{ij} = 2G_0\, de_{ij} - K_{ij}\, dg \tag{2.11}$$

where $dg = (\partial g/\partial e_{ij})\, de_{ij}$. The expression K_{ij} can be shown to be

$$K_{ij} = q(E_2, E_3^2)E_2 e_{ik}e_{kj} + p(E_2, E_3^2)e_{ij} + v(E_2, E_3^2)E_3\delta_{ij}$$

where the quantities E_2 and E_3 are contracted tensor products defined as

$$E_2 = \tfrac{1}{2}e_{mn}e_{mn} \quad \text{and} \quad E_3 = \tfrac{1}{3}e_{mn}e_{np}e_{pm}. \tag{2.12}$$

This representation of a series of odd powered strain deviations is based on arguments similar to those given by Reiner [7], Prager [8], Handelman, Lin, and Prager [2] for higher powers of the stress deviation. The functions $q(E_2, E_3^2)$, $p(E_2, E_3^2)$, and $v(E_2, E_3^2)$ are not entirely arbitrary but subject to the condition

[2] Since a state of hydrostatic pressure produces no deformation in an incompressible material, the only deformation is due to change of shape which is given by the stress deviation.

that $K_{ii} = 0$ (see Eq. (2.11)). For non-vanishing third invariant E_3, $v = -2E_2q/3$, thus K_{ij} becomes

$$K_{ij} = p(E_2, E_3^2)e_{ij} + q(E_2, E_3^2)E_3e_{ik}e_{kj} - 2q(E_2, E_3^2)E_2E_3/3\delta_{ij} \quad (2.13)$$

where p and q are positive definite functions. A more convenient form can be obtained by noting that

$$\partial E_2/\partial e_{ij} = e_{ij} \quad \text{and} \quad \partial E_3/\partial e_{ij} = e_{ik}e_{kj}. \quad (2.14)$$

Assuming sufficiently well behaved functions p and q, there exists an integrating factor $1/G(E_2, E_3^2)$ such that

$$(p/G)\partial E_2/\partial e_{ij} + (q/2G)\partial E_3^2/\partial e_{ij} = \partial g/\partial e_{ij}. \quad (2.15)$$

Here g is a function of E_2 and E_3^2 with the property that

$$\partial g/\partial E_2 = p/G \quad \text{and} \quad \partial g/\partial E_3^2 = q/2G. \quad (2.16)$$

Representation (2.13) becomes

$$K_{ij} = G\partial g/\partial e_{ij} - (4G/3)E_2E_3\partial g/\partial E_3^2\delta_{ij}. \quad (2.17)$$

In a similar manner the differential of a series of strain deviation invariants of even order leads to a differential of a positive definite function in E_2 and E_3^2. It is convenient to use the differential of $g(E_2, E_3^2)$ introduced in Eq. (2.17) as the loading-unloading criterion with the property that an infinitesimal change of strain deviation constitutes loading, unloading, or a neutral change of strain depending on whether dg is positive, negative, or zero. While the vanishing of $d(\epsilon^2/2 + \epsilon^4/4 + \cdots)$ for simple tension has no physical significance, the vanishing of dg for combined states of stress and strain is possible. Neutral change of strain can be regarded as a limiting case of either loading or unloading, and so for that reason stress-strain relations for both should give identical results when $dg = 0$. Thus for combined states of stress and strain, the governing stress-strain relations for the stress theory of plastic flow become

$$ds_{ij} = 2G_0\,de_{ij} - G\{\partial g/\partial e_{ij} - (4E_2E_3/3)\partial g/\partial E_3^2\delta_{ij}\}\,dg \quad \text{for} \quad dg > 0 \quad (2.18)$$

$$ds_{ij} = 2G_0\,de_{ij} \quad \text{for} \quad dg \leqq 0. \quad (2.19)$$

3. Inverted form of the stress-strain relations. For the purpose of finding the inverted form of the stress-strain relation (2.18), it is convenient to let

$$C_{ij} = \partial g/\partial e_{ij} - (4E_2E_3/3)\partial g/\partial E_2^2\delta_{ij} \quad (3.1)$$

and note that

$$dg = (\partial g/\partial e_{ij})\,de_{ij} = C_{ij}\,de_{ij}.$$

Hence (2.18) becomes

$$ds_{ij} = 2G_0\,de_{ij} - GC_{ij}C_{rs}\,de_{rs} \quad \text{for} \quad C_{rs}\,de_{rs} > 0. \quad (3.2)$$

Following the procedure of Greenberg [9] or Hill [10] for strain theories of plastic

flow, Eq. (3.2) can be inverted. Multiplying for C_{ij} gives

$$\begin{aligned} C_{ij}\,ds_{ij} &= 2G_0 C_{ij}\,de_{ij} - GC_{ij}C_{ij}C_{rs}\,de_{rs} \\ &= (2G_0 - GC_{ij}C_{ij})C_{rs}\,de_{rs}\,. \end{aligned} \tag{3.3}$$

The contracted product $C_{rs}\,de_{rs}$ can be evaluated and substituted in Eq. (3.2), resulting in the inverted form of the stress-strain relation, namely

$$de_{ij} = (ds_{ij} + G\{2G_0 - GC_{rs}C_{rs}\}^{-1}C_{ij}C_{kl}\,ds_{kl})/2G_0 \tag{3.4}$$

with a loading condition $(2G_0 - GC_{rs}C_{rs})^{-1}C_{ij}\,ds_{ij} > 0$. The inverted form of the stress-strain relation (2.19) is

$$de_{ij} = ds_{ij}/2G_0 \quad \text{for} \quad (2G_0 - GC_{rs}C_{rs})^{-1}C_{ij}\,ds_{ij} \leqq 0. \tag{3.5}$$

4. Conditions on functions g and G. These functions are not entirely arbitrary, but must be restricted in order that they serve the purpose for which they were intended in the stress-strain relation (2.18). Firstly, we assume that both functions g and G are positive definite forms in the invariants E_2 and E_3^2, and secondly, that the stress-strain relation (2.18) reduces to (2.2) for simple tension. This condition can be conveniently obtained by considering the single scalar equation resulting from a contracted product of Eq. (2.18) with de_{ij},

$$ds_{ij}\,de_{ij} = 2G_0\,de_{ij}\,de_{ij} - G(dg)^2. \tag{4.1}$$

The non-vanishing stress and strain components for simple tension are $\sigma_{11} = \sigma$, $\epsilon_{11} = \epsilon$, $\epsilon_{22} = \epsilon_{33} = -\frac{1}{2}\epsilon$, from which it follows that $E_2 = 3\epsilon^2/4$ and $E_3 = \epsilon^3/4$. Eq. (4.1) becomes

$$\begin{aligned} d\sigma\,d\epsilon = 3G_0\,d\epsilon^2 - 9G/4\{&(\partial g/\partial E_2)^2\epsilon^2 + 2(\partial g/\partial E_2)(\partial g/\partial E_3^2)E_3\epsilon^3 \\ &+ (\partial g/\partial E_3^2)^2E_3^2\epsilon^4\}\,d\epsilon^2. \end{aligned} \tag{4.2}$$

Thus

$$9G/4\{(\partial g/\partial E_2)^2\epsilon^2 + (\partial g/\partial E_2)(\partial g/\partial E_3^2)\epsilon^6/2 + (\partial g/\partial E_3^2)^2\epsilon^{10}/16\}$$

must reduce to $(3a_1\epsilon^2 + 5a_2\epsilon^4 + \cdots)$ for simple tension, where $\partial g/\partial E_2$ and $\partial g/\partial E_3^2$ are positive definite forms (see Eq. (2.16)).

The application of the Drucker work-hardening condition [11] $d\epsilon^p_{ij}\,d\sigma_{ij} \geqq 0$, where $d\epsilon^p_{ij}$ is the differential of the plastic strain tensor, establishes an additional condition on g and G given by the following inequality

$$0 \leqq G\{(\partial g/\partial E_2)^2E_2 + 6(\partial g/\partial E_2)(\partial g/\partial E_3^2)E_3^2 + 4(\partial g/\partial E_3^2)^2E_2^2E_3^2/3\} < G_0\,. \tag{4.3}$$

For incompressible materials the work-hardening condition becomes $de^p_{ij}\,ds_{ij} \geqq 0$. The second term of Eq. (3.4) represents de^p_{ij} so that

$$G\{2G_0(2G_0 - GC_{rs}C_{rs})\}^{-1}(C_{ij}\,ds_{ij})^2 \geqq 0.$$

Since $G_0 > 0$ and $G \geqq 0$, $(2G_0 - GC_{rs}C_{rs}) > 0$ or

$$0 \leqq GC_{rs}C_{rs} < 2G_0\,.^3 \tag{4.4}$$

[3] The author is indebted to the referee for the derivation of this inequality.

From (3.1)

$$C_{rs}C_{rs} = (\partial g/\partial e_{rs})(\partial g/\partial e_{rs}) - 8(\partial g/\partial E_3^2)(\partial g/\partial e_{rr})E_2E_3/3 + 16(\partial g/\partial E_3^2)^2E_2^2E_3^2/3. \quad (4.5)$$

Since $K_{ii} = GC_{ii} = 0$ (see (2.17) and (3.1)), it follows that

$$\partial g/\partial e_{rr} = 4E_2E_3\partial g/\partial E_3^2 .$$

In addition (see (2.14))

$$\partial g/\partial e_{rs} = (\partial g/\partial E_2)e_{rs} + 2E_3(\partial g/\partial E_3^2)e_{rp}e_{ps} .$$

Applying (2.12) and the contracted product of the identity $e_{rp}e_{ps}e_{sq} = E_2e_{rq} + E_3\delta_{rq}$ with e_{qr} gives

$$(\partial g/\partial e_{rs})(\partial g/\partial e_{rs}) = 2E_2(\partial g/\partial E_2)^2 + 12E_3^2(\partial g/\partial E_2)(\partial g/\partial E_3^2) + 8E_2^2E_3^2(\partial g/\partial E_3^2)^2.$$

Hence

$$C_{rs}C_{rs} = 2E_2(\partial g/\partial E_2)^2 + 12E_3^2(\partial g/\partial E_2)(\partial g/\partial E_3^2) + 8(\partial g/\partial E_3^2)^2E_2^2E_3^2/3$$

which together with (4.4) establishes (4.3).

The experimental determination of the stress-strain relation (2.18) can be done by a series of tests described in [12] and [13]. We merely demand that the conditions imposed on g and G in this section be satisfied.

5. Variational principles. A formulation of boundary value problems in connection with the plastic flow theory of materials exhibiting a gradual transition from the elastic to the plastic state has been given by Prager [14], where the state of stress is known throughout the deformed body, while velocities or rates of change of surface stresses are given on the surface. Boundary value problems of this type are natural for the stress-strain relations of the so-called strain theory of plastic flow. In this section we shall consider a type of boundary value problem more suitable to the stress theory of plastic flow.

Let V be any bounded simply connected closed region (over which the divergence theorem holds) with surface S, and let a differential of the displacement vector du_i be given over a portion of the surface, say S_1, such that the surface integral of the normal component of the differential displacement vector vanishes, and a differential of the surface stresses be given over the remaining portion S_2 such that $dT_i = d\sigma_{ij}n_j$ where n_j are the components of the outer unit normal of S with respect to V. The differential of stress deviation ds_{ij} is determined at every point of V by stress-strain relations (2.18) or (3.2) and (2.19) whenever the state of strain e_{ij} is given throughout V and provided the differential of strain de_{ij} can be determined at each point. It follows that the evaluation of de_{ij} constitutes the major problem here. We proceed as follows.

Let a set of primed differentials of strain deviations de'_{ij} be called admissible if they are derivable from a set of differentials du'_i by means of the differential form of Eq. (2.7) with the property that $du'_{i,i} = 0$ throughout the body and $du'_i = du_i$ on S_1. The corresponding differentials ds'_{ij} for a fixed state of strain are

$$ds'_{ij} = 2G_0\, de'_{ij} - GC_{ij}C_{rs}\, de'_{rs} \quad \text{for} \quad C_{rs}\, de'_{rs} > 0 \tag{5.1}$$

$$ds'_{ij} = 2G_0\, de'_{ij} \quad \text{for} \quad C_{rs}\, de'_{is} \leqq 0. \tag{5.2}$$

The actual differential of strain de_{ij} existing in the body is an admissible one whose corresponding differential of stress deviation ds_{ij} satisfies the equations of equilibrium.

Minimum Principle for Differential of Strain Deviation. For any admissible set of differentials of strain and corresponding differentials of stress deviation given by (5.1) or (5.2), the integral

$$J(de'_{ij}) = \tfrac{1}{2}\int_V de'_{ij}\, ds'_{ij}\, dV - \int_{S_2} du'_i\, dT_i\, dS$$

is an absolute minimum for the actual differentials of stress and strain occurring in the deformed body, i.e.

$$J(de'_{ij}) - J(de_{ij}) \geqq 0$$

where

$$J(de_{ij}) = \tfrac{1}{2}\int_V de_{ij}\, ds_{ij}\, dV - \int_{S_2} du_i\, dT_i\, dS.$$

The boundary condition dT_i is given on S_2. Equality can hold if and only if $de'_{ij} = de_{ij}$. It follows that if a solution de_{ij} of the mixed boundary value problem exists it is unique. The proof depends on showing that $J(de'_{ij})$ is equal to $J(de_{ij})$ plus terms whose sum is positive or zero.

If we turn our attention to stress-strain relations (3.4) and (3.5), we note that for a given state of strain e_{ij} at every point of V, de_{ij} can be determined provided ds_{ij} is known throughout the body. Here the major problem constitutes the evaluation of ds_{ij}. Let us define an admissible set of differential stress deviations ds''_{ij} as one whose corresponding differential stress tensors $d\sigma''_{ij}$ (determined only to within an arbitrary constant hydrostatic pressure) satisfy the equations of equilibrium and the boundary condition on S_2, i.e. $d\sigma''_{ij}n_j = dT''_i = dT_i$ on S_2. The corresponding differentials of strain are

$$de''_{ij} = \{(ds''_{ij} + G[2G_0 - GC_{rs}C_{rs}]^{-1})/2G_0\}C_{ij}C_{kl}\, ds''_{kl} \quad \text{for} \quad C_{kl}\, ds''_{kl} > 0, \tag{5.3}$$

and

$$de''_{ij} = (2G_0)^{-1}\, ds''_{ij} \quad \text{for} \quad C_{kl}\, ds''_{kl} \leqq 0 \tag{5.4}$$

where the loading condition is simplified due to the inequality (4.3). The actual differential of stress deviation ds_{ij} occurring in V is an admissible one whose corresponding de_{ij} satisfies the compatibility equations.

Minimum Principle for Differential of Stress Deviation. For any admissible set ds''_{ij} and de''_{ij} given by (5.3) or (5.4), the integral

$$I(ds''_{ij}) = \tfrac{1}{2}\int_V ds''_{ij}\, de''_{ij}\, dV - \int_{S_1} dT''_i\, du_i\, dS$$

is an absolute minimum for the actual differentials of stress and strain occurring in the deformed body, i.e.

$$I(ds''_{ij}) - I(ds_{ij}) \geqq 0$$

where

$$I(ds_{ij}) = \tfrac{1}{2}\int_V ds_{ij}\, de_{ij}\, dV - \int_{S_1} dT_i\, du_i\, dS.$$

The boundary condition du_i is given on S_1. Equality holds if and only if $ds''_{ij} = ds_{ij}$. It follows that if a solution exists it is unique. Both of these principles can be proved by methods similar to those given by Hill [15] or Greenberg [16], since the stress-strain laws they are concerned with in proving minimum principles have the same algebraic structure as those that are considered here. It follows that inequality (4.3) must be applied in proving the first principle, while in the second the proof is quite straightforward.

6. Concluding remarks. The stress-strain relation for loading given by (2.18), (3.2), or (3.4) is limited to those cases where a single loading is followed by at most one unloading. This restriction is necessary because the behavior of the material under simple tension (or more general states of stress and strain) no longer is given by Fig. 1 on repeated loadings.

It is of interest to note that the stress-strain law of Hodge and Prager [17] for incompressible materials is

$$de_{ij} = (2G_0)^{-1}\, ds_{ij} + HD_{ij}D_{pq}\, ds_{pq}$$

where $D_{ij} = \partial f/\partial\sigma_{ij}$ and H and f are positive definite functions of stress invariants. This form is similar to (3.4). On the other hand the stress-strain relation for the stress theory of plastic flow (3.2) is similar to the inverted form of the Hodge and Prager law [16], i.e. the inverted form of the stress-strain relation for the strain theory of plastic flow.

BIBLIOGRAPHY

1. W. Prager, On isotropic materials with continuous transition from elastic to plastic state, Proc. 5th Internat. Congr. Appl. Mech., Cambridge (Mass.), 234–237 (1938).
2. G. H. Handelman, C. C. Lin, and W. Prager, On the mechanical behavior of metals in the strain-hardening range, Q. Appl. Math. 4, 397–407 (1947).
3. D. Trifan, On the plastic bending of circular plates under uniform transverse loads, Q. Appl. Math. 6, 417–427 (1948).
4. L. Prandtl, Spannungsverteilung in plastischen Korpern, Proc. 1st Internat. Congr. Appl. Mech., Delft, 43–54 (1924).
5. E. Reuss, Berucksichtigung der elastischen Formaenderungen in der Plastizitaetstheorie, Z. angew. Math. Mech. 10, 266–274 (1930).
6. D. Trifan, A new theory of plastic flow, Q. Appl. Math. 7, 201–211 (1949).
7. M. Reiner, A mathematical theory of dilatancy, Am. J. Math. 67, 350–362 (1945).
8. W. Prager, Strain-hardening under combined stresses, J. Appl. Phys. 16, 837–840 (1945).
9. H. J. Greenberg, On the variational principles of plasticity, Monograph A 11-S 4, Grad. Div. of Appl. Math., Brown Univ. (March 1949), p. 80.

10. R. Hill, The mathematical theory of plasticity, Oxford Press, 1950, p. 68.
11. D. C. Drucker, Some implications of work hardening and ideal plasticity, Q. Appl. Math. 7, 411–418 (1949).
12. W. Lode, Der Einfluss der mittleren Hauptspannung auf das Fliessen der Metalle, Forschungsarbeiten a.d. Gebiete d. Ingenieurwesens, No. 303, VDI-Verlag, Berlin (1928).
13. G. I. Taylor and H. Quinney, The plastic distortion of metals, Phil. Trans. Roy. Soc. London (A) 230, 323–362 (1931).
14. W. Prager, Fundamental theorems of a new mathematical theory of plasticity, Duke Math. J. 9, 228–233 (1942).
15. R. Hill, loc. cit., 63–66.
16. H. J. Greenberg, loc. cit. p. 60, p. 78.
17. P. Hodge and W. Prager, A variational principle for plastic materials with strain hardening, J. Math. and Phys., 27, 1–10 (1948).

University of Arizona

(Received May 2, 1955)

ON THE PROBLEM OF BUCKLING OF A HYPERBOLIC PARABOLOIDAL SHELL LOADED BY ITS OWN WEIGHT[1]

By Anthony Ralston

1. Formulation of the Problem. The problem which we wish to solve is that of finding the smallest uniformly distributed load which causes the buckling of a shallow hyperbolic paraboloidal shell which has moment free support and whose edge stiffening members are rigid in the direction of their axes and have negligible bending resistance in planes tangent to the shell.

The equation of the undeflected middle surface of this shell is of the form

$$z = c\frac{x}{a}\frac{y}{b} \tag{1}$$

and the weight of the shell per unit surface area is designated by p_0.

The differential equations of the classical linear stability theory for this problem have been derived by E. Reissner [1]. They are of the form

$$\nabla^2\nabla^2\Phi = Eh\frac{2c}{ab}\frac{\partial^2 w}{\partial x \partial y} \tag{2}$$

$$D\nabla^2\nabla^2 w = -\frac{2c}{ab}\frac{\partial^2\Phi}{\partial x\partial y} + p_0\frac{ab}{c}\frac{\partial^2 w}{\partial x\partial y} \tag{3}$$

where h is the thickness of the shell, E is Young's modulus, w is the deflection of the shell, Φ is a stress function and $D = Eh^3/12(1-\nu^2)$.

The boundary conditions corresponding to the assumed edge restraints are [1]

$$\left.\begin{matrix} x = 0, a \\ y = 0, b \end{matrix}\right\} w = \nabla^2 w = \Phi_{xx} = \Phi_{yy} = 0 \tag{4}$$

the subscripts denoting partial differentiation. To these boundary conditions we add the one further condition that the integral of Φ_{xy} along each of the four edges vanish. That is, we prescribe that the resultant shear force which each of the edge stiffeners may apply to the shell is the same as the resultant shear force in the unbuckled state. Using this condition in conjunction with (4) the following important modification of (4) may be effected

$$\left.\begin{matrix} x = 0, a \\ y = 0, b \end{matrix}\right\} w = \nabla^2 w = \Phi = \nabla^2\Phi = 0 \tag{5}$$

The significance of (5) is firstly that now the boundary conditions for w and Φ

[1] This research was supported by the Office of Naval Research under contracts N5ori60 and N5ori07834 with M.I.T. The author wishes to express his appreciation to Professor Eric Reissner, under whose supervision this work was done, for his many valuable suggestions.

are precisely the same and secondly that they are identical with the boundary conditions for the transverse deflection of a simply supported rectangular flat plate. Our object is to find the minimum load, p_0, which allows us to find non-zero solutions of (2) and (3) which satisfy (5).

Equation (3) without the term in Φ_{xy} is the differential equation for the rectangular flat problem if we identify $p_0\, ab/2c$ with $h\tau_{xy}$, the resultant shearing stress applied to the edges of the plate. Using ω for the transverse deflection and replacing $2h\tau_{xy}$ by λ the differential equation for the flat plate problem is

$$D\nabla^2\nabla^2\omega = \lambda \frac{\partial^2\omega}{\partial x \partial y} \tag{6}$$

and the boundary conditions are, as we stated above, precisely the same as those for w and Φ in (5).

Equation (6) together with its boundary conditions is an eigenvalue problem with solutions ω_n corresponding to eigenvalues λ_n. The smallest eigenvalue (in absolute value) corresponds to the critical shearing stress at which the plate buckles. We will consider the eigenvalues with distinct magnitudes to be enumerated so that

$$|\lambda_1| < |\lambda_2| < |\lambda_3| < \cdots \tag{7}$$

For the shell problem as given by equations (2), (3) and (5) let

$$w = A\omega_n \qquad \Phi = B\omega_n \qquad A, B \text{ constants} \tag{8}$$

Since each ω_n satisfies the same boundary conditions which are to be satisfied by w and Φ, both w and Φ in (8) satisfy (5). Substituting (8) into (2) and (3), using (6) and collecting terms we get

$$\left[\frac{B\lambda_n}{D} - AEh\frac{2c}{ab}\right]\frac{\partial^2\omega_n}{\partial x\partial y} = 0 \tag{9}$$

$$\left[A\left(\lambda_n - \frac{p_0 ab}{c}\right) + B\frac{2c}{ab}\right]\frac{\partial^2\omega_n}{\partial x\partial y} = 0 \tag{10}$$

For a non-zero solution of this system to exist the determinant of the coefficients, A and B, must vanish. Expanding the determinant we find that the eigenvalues of the present problem, that is the admissible values of p_0 for a non-zero solution of (2) and (3), are

$$\frac{p_0^{(n)}ab}{c} = \lambda_n + \frac{DEh}{\lambda_n}\left(\frac{2c}{ab}\right)^2 \tag{11}$$

Since the eigenfunctions ω_n form a complete set of functions we have found a complete set of solutions of (2) and (3) subject to the conditions (5). Thus we have shown that each solution of the flat plate problem is a solution of the shell problem and vice versa. And furthermore we have found the relationship (11) between the eigenvalues of the two problems. To find the load, p_0, which causes the shell to buckle we must then, for given a, b, c and h find the positive λ_n (since p_0 must be positive) which makes p_0 a minimum.

2. Numerical Results. To solve the differential equation (6) subject to the boundary conditions of simple support we assume in the usual manner

$$\omega = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} A_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \tag{12}$$

Substitution in (6) leads to the following infinite set of linear algebraic equations for the constants A_{mn} and the parameter λ

$$\frac{abD}{4}\left(\frac{m^2\pi^2}{a^2} + \frac{n^2\pi^2}{b^2}\right)^2 A_{mn} - 4\lambda \sum_{p=1}^{\infty} \sum_{q=1}^{\infty} \frac{mnpqA_{pq}}{(m^2 - p^2)(n^2 - q^2)} = 0, \tag{13}$$

$$m \pm p, n \pm q \quad \text{odd}, \qquad m, n = 1, 2, \cdots$$

Before solving (13) we make the substitutions

$$\alpha = \frac{a}{b} \qquad \mu = \frac{\pi^4 D\alpha}{16b^2\lambda} \tag{14}$$

in order to nondimensionalize (13). Substitution of (14) into (13) gives

$$\mu\left(\frac{m^2}{\alpha^2} + n^2\right)^2 A_{mn} = \sum_{p=1}^{\infty} \sum_{q=1}^{\infty} \frac{mnpqA_{pq}}{(m^2 - p^2)(n^2 - q^2)} \tag{15}$$

$$m \pm p, n \pm q \quad \text{odd}, \qquad m, n = 1, 2 \cdots$$

The system (15) splits up into two separate systems, one for $m + n$ even and one for $m + n$ odd. In order to solve these two infinite systems we will terminate the summations in (12) at finite values, solve the resulting approximate systems and then see how the solutions converge as we increase the number of terms used. In all the calculations we have performed we have assumed that $\alpha = 1$ (square plate) although the calculations could be carried through equally well for any value of α.

The approximate finite problem was solved on the Whirlwind I computer at M.I.T. for each of the two systems $m + n$ even and odd (hereafter called Cases E and O), starting with only the first two terms of (12) and increasing this to as many as 17 terms in Case E and 18 terms in Case O. The numerical method used to solve the two cases was based on the fact that in each case the matrices of coefficients associated with both sides of (15) are symmetric (if we order the equations properly) and the matrix associated with the left hand side of (15) is positive definite. Computationally then the positive definite matrix was first reduced to its canonical form while simultaneously the elementary operations required to do this were also applied to the second matrix. Then the second matrix was diagonalized.[2] The results of this computation were, if n terms of (12) were used, the n eigenvalues and associated eigenvectors (modes of deflection) of the system of equations (for either Case E or O). Less than n useful eigenvalues were obtained for each approximation, however, due to the fact that to each positive μ satisfying (15) there is also an eigenvalue $-\mu$ (which, using (11) and (14), we see corresponds to a negative p_0) and also there were

[2] This method is more fully described by Corbató in [5].

often single or multiple roots $\mu = 0$ (corresponding to an infinite p_0). The largest absolute value of μ obtained (which corresponds, using (14), to the smallest absolute value of $\lambda = \lambda_1$) was an eigenvalue of Case E and it is this value which corresponds to the buckling stress in the flat plate problem. This has been calculated by various authors (e.g. [2], [3]) with results agreeing very closely to the result we have obtained (see footnote to Table 1). The other eigenvalues of Case E and all of those of Case O correspond to shearing stresses greater than the buckling stress. These are of little interest in the flat plate problem but as equation (11) indicates and as we shall see graphically each positive eigenvalue for both Cases E and O corresponds to a critical load for the shell problem.

Using the first 17 terms of (12) with $m + n$ even 6 positive, non-zero values of μ were obtained for Case E (along with 6 negative and 5 zero eigenvalues). For the two largest eigenvalues the convergence of the approximations to μ using less than 17 terms to the 17 term approximation was very good with each eigenvalue being found with an error of less than 1%. The good convergence was also indicated by the fact that the components of the eigenvectors corresponding to relatively large values of $m + n$ were very small compared to the components corresponding to small $m + n$. The convergence for the third largest eigenvalue was also quite good (less than 5% error). For the fourth eigenvalue the convergence was poor while for the final two 17 terms is not enough to indicate their accuracy. For Case O in the 18 term approximation 8 positive non-zero values of μ were obtained (along with 8 negative and 2 zero eigenvalues). Again the two largest eigenvalues were found very accurately (error less than 1%), the convergence to the third eigenvalue was fair (less than 5% error), the convergence to the fourth eigenvalue was poor while for the last four eigenvalues 18 terms is not enough to indicate their accuracy.

Corresponding to each eigenvalue of both Cases E and O there is an approximation to an eigenvalue of (6) and then using (11) we have an approximation to an eigenvalue of (2) and (3). The closeness of this approximation depends on how well the solutions of (15) converge as discussed above. Using (14) we may express the eigenvalues of (6) in the form

$$\begin{aligned} \lambda_n^{(e)} &= K_n^{(e)} \frac{D}{b^2} && m + n \quad \text{even} \\ \lambda_n^{(0)} &= K_n^{(0)} \frac{D}{b^2} && m + n \quad \text{odd} \end{aligned} \tag{16}$$

where we have used $\alpha = 1$. We further assume that the $\lambda_n^{(e)}$ and $\lambda_n^{(0)}$ are ordered as the λ_n in (7). Values of $K_n^{(e)}$ ($n = 1$ to 6) for the 17 term approximation and values of $K_n^{(0)}$ ($n = 1$ to 6) for the 18 term approximation are found in Table 1.

In order to relate the results of the plate problem to the shell problem we first define $\beta = h/c$ and then define

$$P = \frac{p_0}{\sqrt{DEh}} \left(\frac{ab}{2c}\right)^2 = \frac{p_0 \sqrt{12(1 - \nu^2)}}{4E\beta^2} \frac{b^4}{c^4} \tag{17}$$

TABLE 1

Values of $K_n^{(e)}$ and $K_n^{(o)}$ for equation (16)[3]

n	$K_n^{(e)}$	$K_n^{(o)}$
1	184.3	228.0
2	491.2	530.7
3	652.9	614.6
4	833.3	911.5
5	1628.	1366.
6	1801.	1454.

Then substituting (11) into (17) and using (16) we may write

$$P_n^{(e)} = \frac{K_n^{(e)}\beta}{4\sqrt{12(1-\nu^2)}} + \frac{\sqrt{12(1-\nu^2)}}{K_n^{(e)}\beta}$$
$$P_n^{(0)} = \frac{K_n^{(0)}\beta}{4\sqrt{12(1-\nu^2)}} + \frac{\sqrt{12(1-\nu^2)}}{K_n^{(0)}\beta} \tag{18}$$

Using the values of $K_n^{(e)}$ and $K_n^{(0)}$ in Table 1 values of $P_n^{(e)}$ and $P_n^{(0)}$ as a function of β are plotted in Figure 1 for $n = 1, 2, 3$ (using $\nu = .3$). Since the convergence to the three largest values of μ was quite good the curves in Figure 1 are quite accurate since the three largest values of μ in each case correspond to the three smallest values of $K_n^{(e)}$ and $K_n^{(0)}$. The solid curve determines the critical value of P at which the shell becomes unstable. That is, using (17), it determines the minimum p_0 which enables us to find non-zero solutions of (2) and (3). This curve is made up of portions of each $P_n^{(e)}$ and $P_n^{(0)}$ curve and this would still be true if curves for all the $P_n^{(e)}$ and $P_n^{(0)}$ $(n = 1, 2, \cdots)$ were included in Figure 1. This is because, as we can see from (18), each $P_n^{(e)}$ and $P_n^{(0)}$ curve has a minimum at 1 and no $P_n^{(e)}$ or $P_n^{(0)}$ curve ever goes below 1 so that the minimum point for each curve is on the critical curve.[4] Therefore each of the eigenvalues for Cases E and O is the determining one for the buckling load for some value of β.

To actually find the critical load for a given shell with $a = b$ we determine β and then from Figure 1 find the critical value of P. Then using (17) in the form

$$p_0 = \frac{4E\beta^2 P}{\sqrt{12(1-\nu^2)}}\frac{c^4}{b^4} \tag{19}$$

we can calculate the critical load. Although the critical curve in Figure 1 is only shown for values of β as small as .01 it seems likely that for values of β smaller than this the critical curve will remain very close to the minimum line at 1. This is because all the minima on the $P_n^{(e)}$ and $P_n^{(0)}$ curves for $n > 3$ are

[3] $K_1^{(e)}$ corresponds to the critical shearing stress in the flat plate problem. The constant commonly defined in the literature (see [4, p. 360]) is $k = K_1^{(e)}/2\pi^2$ and the value given for k is 9.34. Using the present $K_1^{(e)}$ we get $k = 9.338$.

[4] We note that this minumum value has been obtained earlier by Reissner [1] by considering a special simple solution of the system (2) and (3).

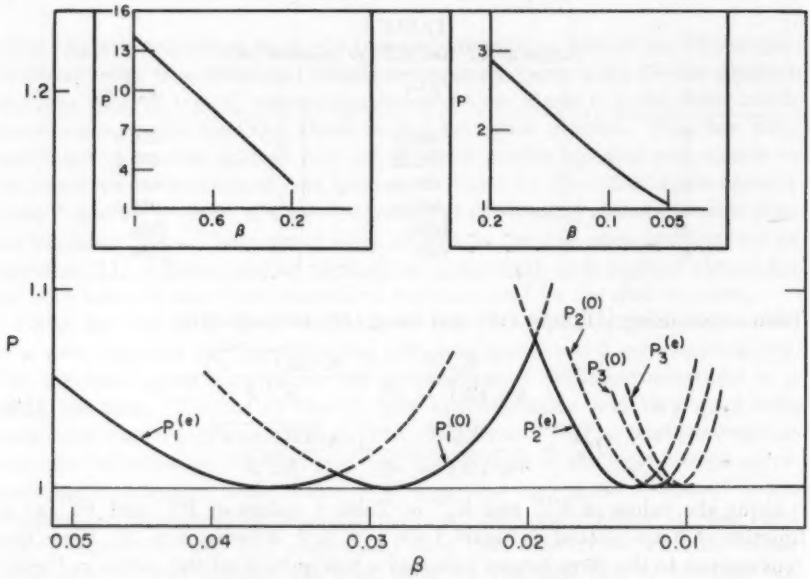


FIG. 1. Buckling load coefficient P as a function of $\beta = h/c$

bunched between β equals .01 and 0. Thus for practical purposes we will be able to take the critical value of P to be equal to 1 for $\beta < .01$. For this reason and because even on the computer the calculations with more than 18 terms of (12) are quite lengthy we have not attempted to determine the critical curve accurately for values of β smaller than .01.

The critical curve in Figure 1 is not so "regular" (e.g. the space between the $P_1^{(0)}$ and $P_2^{(e)}$ portions of the curve is much larger than that between the $P_1^{(e)}$ and $P_1^{(0)}$ portions) as in the case of similar curves for other problems in elasticity (e.g. the curve used to find the critical stress in the problem of a simply supported rectangular flat plate uniformly compressed in one direction [4, p. 330]). This is because Figure 1 is made up of curves derived from the solutions of two separate sets of equations, Cases E, and O, and because, as we stated above, the curves will become increasingly bunched together as β decreases.

The results obtained here may of course be extended to hyperbolic paraboloidal shells where a is not equal to b. But for each value of α a complete new calculation using (15) would have to be made and then a new curve analogous to Figure 1 would have to be plotted. Alternatively if (15) was solved for a sufficient number of values of a/b, curves could be plotted with a/b as abscissa for a constant value of h/c.

REFERENCES

1. E. REISSNER, On Some Aspects of the Theory of Thin Elastic Shells, Journal of the Boston Society of Civil Engineers, **42**, 100 (1955)

2. E. Seydel, Über das Ausbeulen von rechteckigen, isotropen oder anisotropen Platten bei Schubbeanspruchung, Ingenieur-Archiv, **4**, 169 (1933)
3. S. Bergmann and H. Reissner, Neuere Probleme aus der Flugzeugstatik, Z. Flugtech. Motorluftsch., **23**, 6 (1932).
4. S. Timoshenko, Theory of Elastic Stability, McGraw-Hill (1936).
5. F. J. Corbató, Programs for Solving Secular Equations, Digital Computer Laboratory Memorandum DCL-58, Mass. Inst. of Tech. (1955).

Massachusetts Institute of Technology

(Received October 24, 1955)

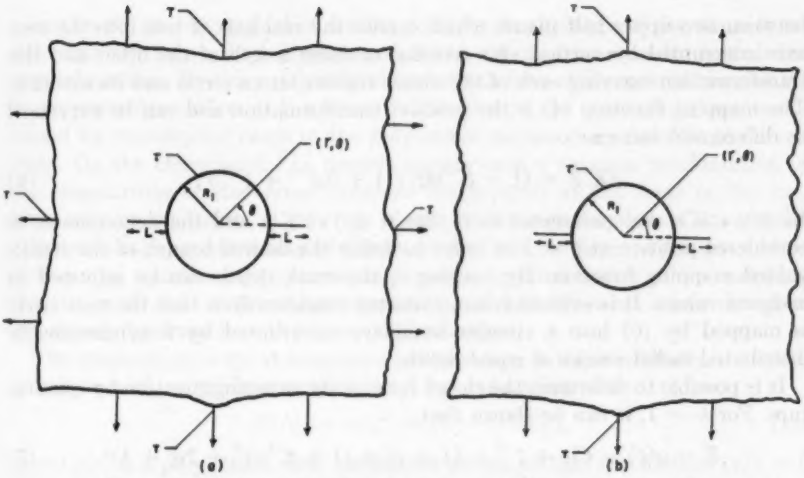


FIG. 1. Geometry and loading for the case of two cracks, ($k = 2$)

and

$$\sigma_y - \sigma_x + 2i\tau_{xy} = 2[\overline{Z}\varphi''(Z) + \psi'(Z)], \tag{3}$$

where the prime notation denotes differentiation with respect to Z and the bars denote complex conjugates.

It is convenient for the purpose of enforcing boundary conditions to introduce an auxiliary complex plane, the ζ-plane, such that the unit circle, $\zeta = \sigma = e^{i\varphi}$ (where φ is used here to denote angular measure in the ζ-plane), and its exterior are mapped into τ and its exterior, respectively, by the analytic function

$$Z = \omega(\zeta). \tag{4}$$

The stress functions $\varphi(Z)$ and $\psi(Z)$ will be considered as functions of the parameter ζ. The necessity for introducing considerable new notation can be avoided by designating $\varphi(Z) = \varphi[\omega(\zeta)]$ as $\varphi(\zeta)$, etc., which leads to such relationships as $\varphi'(Z) = \varphi'(\zeta)/\omega'(\zeta)$, etc. The analyticity of $\varphi(\zeta)$ and $\psi(\zeta)$ which follows from their definitions contributes significantly to the advantage gained by using the complex variable method.

The mathematical analysis of the problem requires the determination of the functions $\varphi(\zeta)$ and $\psi(\zeta)$ which are analytic for $|\zeta| > 1$ (with the exception of the point at infinity) and lead to the proper loading conditions at infinity and on the internal boundary τ. The forms of $\varphi(\zeta)$ and $\psi(\zeta)$ can be chosen a priori to yield the proper loading conditions at infinity. The condition that τ be load-free can be written as

$$\varphi(\sigma) + \omega(\sigma)\overline{\varphi'(\sigma)}/\overline{\omega'(\sigma)} + \overline{\psi(\sigma)} = 0 \tag{5}$$

3. Description of the Mapping Function. The mapping function can be determined in a fairly easy manner by considering separately the transformation

ANALYSIS OF AN INFINITE PLATE CONTAINING RADIAL CRACKS ORIGINATING AT THE BOUNDARY OF AN INTERNAL CIRCULAR HOLE

By O. L. Bowie

1. Introduction. Considerable advance has been made in recent years in the application of energy type of theories in determining the influence of cracks in the specimen geometry on the strength of the specimen. In order to apply theories such as that developed by A. A. Griffith [1]*, it is necessary to calculate the elastic strain energy of the system. Although only the boundary stresses and displacements are actually necessary for this calculation, one must nevertheless formally solve the problem as a whole to obtain this information.

We shall consider the solution of the class of plane problems in elasticity corresponding to a distribution of radial cracks, equal and finite in length, originating at the boundary surface of a circular hole in an infinite plate under the two load systems shown in Figure 1. The geometry of the internal boundary, τ, can be conveniently described by considering the plate as the complex Z-plane where $Z = x + iy = re^{i\theta}$. Then, if the center of the hole is chosen as $Z = 0$, we specify that radial cracks of equal length, L, lie along $\theta = 0, 2\pi/k, \cdots, (k-1)\,2\pi/k$ where $k \geqq 1$ is an integer which specifies the number of cracks.

In addition to the plane stress applications of the case of uniform (all-around) tension at infinity, the analysis by a slight modification enables us to study the plane strain problem corresponding to internal pressure acting in hollow cylinders of large wall thickness with longitudinal cracks originating along the inside surface. Indeed, this latter problem is solved by a modification of the manner in which the elastic constants enter in the displacement relations and by the superposition of the simple solution corresponding to uniform hydrostatic pressure.

2. Stress Problem Formulated in Terms of Complex Variables. Due to the irregular geometry of the internal boundary, it would appear that the problem described above can be most conveniently handled by the complex variable method of Mushelisvili [2]. This method depends upon the representation of the well-known Airy's stress function, $U(x, y)$, in terms of two analytic functions of the complex variable, Z, namely, $\varphi(Z)$ and $\psi(Z)$, where

$$U(x, y) = \mathrm{Re}\left[\bar{Z}\varphi(Z) + \int^{Z}\psi(Z)\,dZ\right]. \tag{1}$$

With this representation, the stress components in rectangular coordinates can be written as

$$\sigma_y + \sigma_x = 2[\varphi'(Z) + \overline{\varphi'(Z)}] = 4\,\mathrm{Re}[\varphi'(Z)] \tag{2}$$

* Numbers in brackets refer to the references at the end of the paper.

In this paper, the method of dealing with the singularities of the stress functions will depend on the representation of the mapping function by polynomial approximations. By using polynomial approximations of the internal boundary, an accurate description of the stress distribution at the crack roots can be obtained by introducing cusps in the polynomial mappings to describe the crack roots. On the other hand, the process circumvents a rigorous consideration of the singularities of the stress functions which occur at the roots of (9), i.e., physically at the points of juncture of the cracks and the circle. Convergence of polynomial approximations of the exact problem can be considered a matter of choosing a sufficiently accurate polynomial approximation of the mapping function such that a closer approximation will not effect an improvement in the specified accuracy of the desired information.

The existence of cusps at locations corresponding to the crack roots is ensured in polynomial mapping approximations by demanding that

$$dZ/d\zeta = \omega'(\zeta) = (1 - \zeta^{-k})g(\zeta), \tag{10}$$

where $g(\zeta)$ is a polynomial with coefficients chosen so that the roots of $g(\zeta) = 0$ fall inside the unit circle. Due to the convergence of (8) at all but a finite number of points on the unit circle, suitable polynomial approximations,

$$Z = \omega(\zeta) = C\left[\zeta + \sum_{n=1}^{N} \epsilon_n \zeta^{1-kn}\right], \tag{11}$$

can be obtained by setting $\epsilon_n \cong A_n$; modifications of the ϵ_n's being made to satisfy (10).

5. "All-Around" Tension at Infinity. Consider now the case of uniform tension at infinity illustrated in Figure 1a. It can easily be shown[2] that the loading condition $\sigma_x = \sigma_y = T$ on $|Z| = R$ where R is very large is satisfied by choosing $\varphi(\zeta)$ and $\psi(\zeta)$ such that they approach $CT\zeta/2$ and $CT\gamma_0\zeta^{-1}$, respectively, for large $|\zeta|$. Therefore, let us assume that $\varphi(\zeta)$ is a polynomial of the form

$$\varphi(\zeta) = CT\left[\zeta/2 + \sum_{n=1}^{N} \alpha_n \zeta^{1-kn}\right]. \tag{12}$$

Such indeed[3] would be the form of $\varphi(\zeta)$ were there no boundary irregularities implied by the mapping function (11).

Next, we write the boundary condition (5) as

$$\omega'(\sigma)\psi(\sigma) = -\omega'(\sigma)\overline{\varphi(\sigma)} - \overline{\omega(\sigma)}\varphi'(\sigma). \tag{13}$$

The function $\omega'(\zeta)\psi(\zeta)$ is analytic exterior to the unit circle and, with $\varphi(\zeta)$ assumed as (12), is given as a continuous function on the unit circle by (13). Thus, if the coefficients α_n can be chosen so that the coefficients of all positive powers of ζ in the Laurent expansion of

$$\omega'(\zeta)\varphi(1/\zeta) + \omega(1/\zeta)\varphi'(\zeta) \tag{14}$$

[2] A detailed discussion of the stress functions for infinite regions can be found in [4].

[3] E.g., the case for $k = 2, N = 2$ has been discussed by Morkovin [5] in some detail.

between two upper half planes which carries the real axis of one into the real axis interrupted by vertical slits (cracks) of finite length of the other and the transformation carrying each of the above regions into a circle and its exterior. The mapping function (4) is the product transformation and can be expressed in differential form as

$$dZ/Z = (1 - \zeta^{-k})d\zeta/\zeta(1 + 2\epsilon\zeta^{-k} + \zeta^{-2k})^{\frac{1}{2}}. \tag{6}$$

In (6), ϵ is a real parameter such that $0 \leqq |\epsilon| \leqq 1$ and the denominator is considered positive at $\zeta = 1$ in order to define the desired branch of the multivalued mapping function. By varying ϵ, the crack depth can be adjusted to assigned values. It is evident from symmetry considerations that the unit circle is mapped by (6) into a circular boundary interrupted by k symmetrically distributed radial cracks of equal depth.

It is possible to determine the closed form of the mapping function by quadrature. For $k = 1$, it can be shown that

$$Z = \omega(\zeta) = C[\zeta + \zeta^{-1} + (1 + \epsilon) + (1 + \zeta^{-1})(\zeta^2 + 2\epsilon\zeta + 1)^{\frac{1}{2}}, \tag{7}$$

where C is a constant of proportionality. The closed form of the mapping function increases in complexity with larger integers, k.

For purposes of the subsequent stress analysis, it is desirable to find a series representation of (6) converging on and exterior to the unit circle. The form of such a series is evidently

$$Z = \omega(\zeta) = C\left[\zeta + \sum_{n=1}^{\infty} A_n \zeta^{1-kn}\right], \tag{8}$$

where the A_n's are real. The A_n's may be obtained numerically from simple recursive formulae determined by expanding both sides of (6) in series form and equating coefficients of equal powers of ζ. The convergence of (8) on and exterior to the unit circle can be studied by examining the coefficients A_n. It can be shown[1] that $\lim_{n\to\infty} A_n = 0$; thus, using a well-known theorem found in [3], the series (8) converges exterior to the unit circle and at all points on the unit circle except at the roots of

$$\zeta^{2k} + 2\epsilon\zeta^k + 1 = 0. \tag{9}$$

4. Polynomial Approximation of the Mapping Function. A mathematical difficulty common to most problems in the stress analysis of regions involving irregularities in boundary contours is the determination a priori of the forms of singularities of the stress function associated with the irregularity. Often, the issue is avoided by assuming that the singularities are of such a nature that the stress functions can still be expanded in an infinite series, converging in the region of the problem at all but the singular points of the boundary. Although results which are undoubtedly valid are usually obtained when this argument is used, it is virtually impossible to verify rigorously the convergence assumed.

[1] The proof is particularly simple for $k = 1$. From (7), it can be seen that the A_n's behave essentially as the Legendre polynomials. The proof is more difficult for $k \geqq 2$.

vanish, we can determine $\psi(\zeta)$ explicitly. By multiplying both sides of (13) by $1/2\pi i(\sigma - \zeta)$ and integrating around the unit circle, we obtain by a well-known theorem[4]

$$\omega'(\zeta)\psi(\zeta) = -\omega'(\zeta)\varphi(1/\zeta) - \omega(1/\zeta)\varphi'(\zeta) \tag{15}$$

It is interesting to note that the cusp roots are reflected by singularities in $\psi(\zeta)$ in the form of simple poles.

To complete the analysis it is necessary to verify that the coefficients α_n can be determined to meet the condition set forth above. If the coefficients of all positive powers of ζ in the Laurent expansion of (14) are equated to zero, the following system of linear simultaneous equations results:

$$\alpha_p + \sum_{n=1}^{N-p} \alpha_{p+n}\epsilon_n(1 - nk) + \sum_{n=1}^{N-p} \epsilon_{p+n}\alpha_n(1 - nk) + \epsilon_p/2 = 0, \qquad p = 1, 2, \cdots, N. \tag{16}$$

The conditions can therefore be met by the solution of (16) for the N coefficients, α_n, $n = 1, 2, \cdots, N$.

It is evident that the state of stress is unaffected by the addition of a linear expression of the form $DiZ + a + ib$ to $\varphi(Z)$ or a complex constant $\alpha + i\beta$ to $\psi(Z)$. Thus, for $k = 1$, the quantity α_1 determined by (16) plays no role in the actual stresses. Similarly, identical satisfaction of the boundary condition (5) can be realized by proper adjustments of the coefficients of the above arbitrary additions.

6. Tension in One Direction at Infinity. For the case of tension in one direction at infinity (illustrated in Figure 1b), the analysis can be carried out in a manner similar to that of the preceding section. It can be shown that the loading condition at infinity, $\sigma y = T$, is satisfied by requiring the stress functions $\varphi(\zeta)$ and $\psi(\zeta)$ to approach $CT\zeta/4$ and $CT\zeta/2$, respectively, for large $|\zeta|$. The stress function $\varphi(\zeta)$ is again assumed to have the form of a polynomial

$$\varphi(\zeta) = CT\left[\zeta/4 + \sum_{n=1}^{kN} \alpha_n \zeta^{1-n}\right]. \tag{17}$$

By an argument similar to that of the preceding section, it can again be shown that

$$\omega'(\zeta)\psi(\zeta) = -\omega'(\zeta)\varphi(1/\zeta) - \omega(1/\zeta)\varphi'(\zeta) \tag{18}$$

provided that the coefficients α_n are chosen so that

$$-\omega'(\zeta)\varphi(1/\zeta) - \omega(1/\zeta)\varphi'(\zeta) \approx C^2T\zeta/2, \quad \text{for large} \quad |\zeta|. \tag{19}$$

Due to the relative lack of symmetry in this case, it is difficult to present the linear systems of simultaneous equations for the determination of the α_n's in compact form for arbitrary integers k. Therefore, only the systems for the single

[4] Reference [4], p. 145.

crack ($k = 1$) and the two crack ($k = 2$) cases will be explicitly formulated.

For $k = 1$,

$$\alpha_p + \sum_{n=1}^{N-p} \alpha_{n+p}\epsilon_n(1-n) + \sum_{n=1}^{N-p} \epsilon_{n+p}\alpha_n(1-n) + \epsilon_p/4 = \begin{cases} 0, & p \neq 2 \\ -\frac{1}{2}, & p = 2 \end{cases} \quad (20)$$

$$p = 1, 2, \cdots, N$$

For $k = 2$,

$$\alpha_{2p} + \sum_{n=1}^{N-p} \alpha_{2(n+p)}\epsilon_n(1-2n) + \sum_{n=1}^{N-p} \epsilon_{n+p}\alpha_{2n}(1-2n) + \epsilon_p/4 = \begin{cases} 0, p > 1 \\ -1/2, p = 1 \end{cases} \quad (21)$$

$$\alpha_{2p-1} = 0, \qquad p = 1, 2, \cdots, N,$$

7. Stability Condition of the Griffith Theory for Brittle Failure. An application of the preceding results is the determination of the critical applied load for which the radial cracks begin to spread. A purely elastic criterion is found in the Griffith theory of the rupture of brittle materials [1]. The Griffith hypothesis, when applied to the present problem, leads to the following stability condition:

$$\frac{dV}{dL}\,dL = -2khGdL, \quad (22)$$

where

$V = V_e - V_0$ = reduction of the potential energy due to the presence of the radial cracks,
V_e = potential energy of the system with radial cracks,
V_0 = potential energy of the system without cracks,
k = number of radial cracks,
h = thickness of the specimen (axially),
G = surface tension per unit area of the crack surface,
L = length of the radial crack(s),
d = notation for differential.

The critical applied load can be determined from (22) once the reduction of potential energy, V, has been found as a function of L.

8. Calculation of the Reduction in Potential Energy, V, for the Case of "all-around" Tension at Infinity. In order to compare the difference in potential energies in the physical systems with and without cracks, it is convenient and more motivating to refer to the exact geometry rather than polynomial approximations. Accordingly, the calculation of V first will be carried out in terms of the stress functions as referred to the original Z-coordinate system and the exact boundary shape.

For the case of uniform tension at infinity, the potential energy of the system with cracks, V_c, under the assumption of plane stress, reduces to the calculation of

$$V_c = \lim_{R\to\infty}\left\{-\tfrac{1}{2}h\int_0^{2\pi}(\sigma_r U_r + \tau_{r\theta}U_\theta)R\,d\theta\right\} \tag{23}$$

where σ_r, $\tau_{r\theta}$, U_r and U_θ are stress and displacement components in polar coordinates. The integrand in (23) can be expressed in terms of the original stress functions $\varphi(Z)$ and $\psi(Z)$ defined by (2) and (3) from the relations

$$\sigma_r - i\tau_{r\theta} = \varphi'(+Z)\overline{\varphi'(Z)} - e^{2i\theta}[\overline{Z}\varphi''(Z) + \psi'(Z)] \tag{24}$$

and

$$2\mu(U_r + iU_\theta) = e^{-i\theta}[\eta\varphi(Z) - Z\overline{\varphi'(Z)} - \overline{\psi(Z)}], \tag{25}$$

where, in terms of Young's Modulus, E, and Poisson's ratio, ν,

$$\mu = E/2(1+\nu) \tag{26}$$

and

$$\eta = \begin{cases}(3-\nu)/(1+\nu), & \text{for plane stress}\\ 3-4\nu, & \text{for plane strain.}\end{cases} \tag{27}$$

The stress functions $\varphi(Z)$ and $\psi(Z)$ are certainly analytic exterior to τ except for the point at infinity. Therefore, for the case of uniform tension at infinity the stress functions may be expanded in series form,

$$\varphi(Z) = T[Z/2 + a_1Z^{1-k} + \cdots] \tag{28}$$

and

$$\psi(Z) = T[b_0Z^{-1} + b_1Z^{-1-k} + \cdots]. \tag{29}$$

The expansions (28) and (29), valid for large R, can be substituted into (23) to determine V_c. It follows, after some algebra, that

$$V_c = \lim_{R\to\infty}\left\{-\frac{\pi T^2h}{E}[(1-\nu)R^2 - 2\nu b_0]\right\}. \tag{30}$$

In order to obtain the potential energy, V_0, for the uncracked plate, the stress analysis for a concentric ring with inner radius R_1 and outer radius R loaded on the outer boundary by the same applied load (as the system with cracks) was carried out. This analysis was straightforward and the details will not be included here. The results lead to the value of the potential energy of the uncracked plate.

$$V_0 = \lim_{R\to\infty}\left\{-\frac{\pi T^2h(R^2+b_0)^2[(1-\nu)R^2 + 2\nu R_1^2]}{E(R^2-R_1^2)^2}\right\} \tag{31}$$

In the limit, therefore, the reduction in potential energy for the case of "all-

around" tension at infinity for the plane stress assumption is given by

$$V = 2\pi T^2 h(b_0 + R_1^2)/E. \tag{32}$$

We now must interpret this result in terms of the previous analysis based on polynomial approximations of the problem. The stress function $\psi(\zeta)$ given by (15) for polynomial approximations of the geometry can be expanded in a series of the form

$$\psi(\zeta) = CT[\gamma_0' + \gamma_0\zeta^{-1} + \gamma_1\zeta^{-1-h} + \cdots]. \tag{33}$$

On the other hand, a series expansion of $\psi(\zeta)$ for the exact geometry which is certainly valid for $|\zeta| > 1$ can be found by substituting the exact mapping (6) into (29). If we assume that the polynomial approximation can be made to converge to the exact solution, we find by the comparison of the coefficients of ζ^{-1} in the expansions of $\psi(\zeta)$ that

$$C^2\gamma_0 \to b_0 \tag{34}$$

The hitherto unspecified constant C occurring in the mapping function will now be chosen so that the radius of the circular hole in the physical plane is the unit of length. Thus, if $\sigma = \sigma_1$ is that point on the unit circle in the ζ-plane which corresponds to the junction of the crack and the circle in the Z-plane, then C is chosen so that

$$R_1 = |\omega(\sigma_1)| = 1. \tag{35}$$

Thus, if a sufficiently accurate polynomial mapping is taken, we can write (32) within an arbitrarily specified accuracy as

$$V = 2\pi T^2 h\{\gamma_0/[\omega(\sigma_1)/C]^2 + 1\}/E. \tag{36}$$

The reduction in potential energy can be determined as a function of the crack length, L, by noting that L, measured in units of the radius of the circular hole, can be written as

$$L = \omega(1)/\omega(\sigma_1) - 1. \tag{37}$$

Thus, for the case of plane stress and "all-around" tension at infinity,

$$V = 2\pi T^2 h f(L)/E, \tag{38}$$

where,

$$f(L) = \gamma_0[(L + 1)/\omega(1)/C]^2 + 1. \tag{39}$$

In (39), it should be noted that both γ_0 and $\omega(1)/C$ are functions of L.

For plane strain, the second form of η in (27) must be used for the calculation of the displacements. Otherwise, the calculation is identical with that for the case of plane stress. The reduction in potential energy for the case of plane strain is given by

$$V = 2\pi T^2 h(1 - \nu^2)f(L)/E \tag{40}$$

where $f(L)$ is given again by (39).

9. The Reduction in Potential Energy for the Case of Simple Tension at Infinity. The calculation of V for the case of simple tension at infinity, $\sigma y = T$, can be carried out in a manner similar to the preceding section. For the case of plane stress, we find

$$V = 2\pi T^2 h g(L)/E, \tag{41}$$

where

$$g(L) = (\gamma_0 + 2\alpha_2 - A)[(L+1)/\omega(1)/C]^2/2 + \tfrac{3}{4}. \tag{42}$$

In (42), the quantities γ_0, α_2, and A are defined by

$$\begin{aligned} CT\gamma_0 &= \text{coefficient of } \zeta^{-1} \text{ in the Laurent expansion of } \psi(\zeta), \\ \alpha_2 &= \text{coefficient of } \zeta^{-1} \text{ in the Laurent expansion of } \varphi(\zeta), \\ CA &= \text{coefficient of } \zeta^{-1} \text{ in the Laurent expansion of } \omega(\zeta). \end{aligned} \tag{43}$$

10. The critical load according to the Griffith Hypothesis. The critical value of the applied load, T_c, can now be calculated for the cases considered by substituting (38), (40), or (41) into the stability condition, (22). The results can be summarized as follows:
For plane stress,

$$T_c = [-kEG/\pi f'(L)]^{\frac{1}{2}} \quad \text{for uniform tension at infinity,} \tag{44}$$

and

$$T_c = [-kEG/\pi g'(L)]^{\frac{1}{2}} \quad \text{for simple tension at infinity.} \tag{45}$$

For plane strain,

$$T_c = [-kEG/\pi(1-\nu^2)f'(L)]^{\frac{1}{2}} \quad \text{for uniform tension at infinity.} \tag{46}$$

According to the Griffith hypothesis, the crack(s) become unstable if $T \geqq T_c$.

The plane strain criterion (46) is intended for the study of radial cracks in cylinders under internal pressure, P. Such a load system is obtained simply by superimposing a hydrostatic pressure on the load system corresponding to the uniform tension case for $T = P$. Since it can be shown that the superposition of a hydrostatic stress state does not affect the critical stress as calculated above, the critical internal pressure, P_c, is also given by (46).

11. Numerical Results for the Case of "All-Around" Tension at Infinity." The analysis carried out in the previous sections was numerically evaluated for the cases of a single crack ($k = 1$) and two cracks ($k = 2$). In particular, the stress analysis was carried out for nine values of the crack length-determining parameter, ϵ. From these results, the function $f(L)$ was found in tabular form. These data, in turn, were numerically differentiated with respect to L to find $f'(L)$. The critical stresses according to (44) or (46) were then determined.

For example, when $k = 2$ the coefficient γ_0 in (39) is given by

$$\gamma_0 = -\left[1 + 2\sum_{n=1}^{N} \epsilon_n \alpha_n (1 - 2n)\right]. \tag{47}$$

It was found that γ_0 can be obtained to three significant figure accuracy for the crack lengths chosen if an average of thirty terms of the polynomial approximation of the mapping function is retained. Actually, only a few terms of the polynomial are required for moderately large crack lengths, whereas the convergence is poor for small lengths. Illustrative results for the case of $k = 2$ are listed in Table 1.

TABLE 1

ϵ	L	$\omega(1)/c$	γ_0	$f(L)$
−1.000	0.000	1.000	−1.000	0.000
−0.866	0.303	1.259	−1.075	−0.152
−0.707	0.497	1.383	−1.160	−0.359
−0.500	0.732	1.500	−1.267	−0.689
0.000	1.414	1.707	−1.509	−2.018
+0.500	2.732	1.866	−1.752	−6.008
+0.707	4.027	1.924	−1.852	−11.640
+0.866	6.596	1.966	−1.933	−27.860
+1.000	∞	2.000	−2.000	$-\infty$

For large values of L, the solution approaches the form corresponding to the case of a single crack of length $2(1 + L)$ in an infinite plate. Thus, for large L

$$f(L) \approx -(L + 1)^2/2 + 1. \tag{48}$$

Similar computations were made for the case of a single crack ($k = 1$). The form of $f(L)$ for large L in this case was found to be

$$f(L) \approx -(L + 2)^2/8 + 1. \tag{49}$$

The variation of the critical load, T_c, was then found by differentiating $f(L)$ numerically and substituting into (44) or (46). For the case of plane strain, the variation of the critical load with crack length is shown in Figure 2.

12. Numerical Results for the Case of Simple Tension at Infinity. In a similar manner the cases for $k = 1$ and $k = 2$ were evaluated for the loading $\sigma_y = T$ at infinity. For example, to calculate $g(L)$ in (42) the quantity γ_0 for $k = 2$ is determined from

$$\gamma_0 = \epsilon_1/2 - \left[1/2 + 2\sum_{n=1}^{N} \epsilon_n \alpha_{2n} (1 - 2n)\right] \tag{50}$$

Illustrative data for $k = 2$ is shown in Table 2.

For large crack lengths $g(L)$ approaches the respective forms of $f(L)$ in (48) and (49), which is consistent with the well-known result that a single crack in an infinite sheet is unaffected by a tension in the direction of the crack. The critical stresses, determined by substituting the function $g'(L)$ into (45), are shown as a function of L in Figure 3.

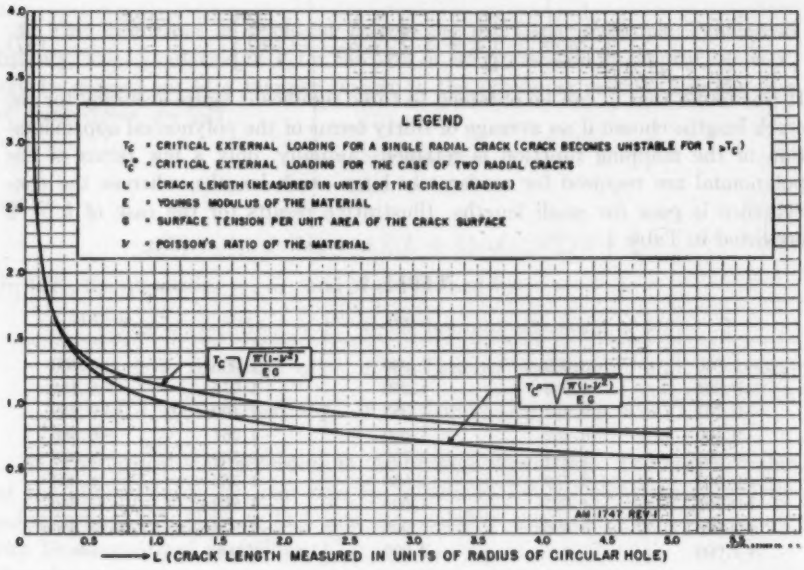

LEGEND
T_c = CRITICAL EXTERNAL LOADING FOR A SINGLE RADIAL CRACK (CRACK BECOMES UNSTABLE FOR $T > T_c$)
T_c^* = CRITICAL EXTERNAL LOADING FOR THE CASE OF TWO RADIAL CRACKS
L = CRACK LENGTH (MEASURED IN UNITS OF THE CIRCLE RADIUS)
E = YOUNGS MODULUS OF THE MATERIAL
G = SURFACE TENSION PER UNIT AREA OF THE CRACK SURFACE
ν = POISSON'S RATIO OF THE MATERIAL
$T_c\sqrt{\frac{\pi(1-\nu^2)}{EG}}$
$T_c^*\sqrt{\frac{\pi(1-\nu^2)}{EG}}$
AM 1747 REV I
L (CRACK LENGTH MEASURED IN UNITS OF RADIUS OF CIRCULAR HOLE)

FIG. 2. Variation of the critical load with crack length for the case of "all-around" tension at infinity.

TABLE 2

ϵ	α_1	α_2	γ_0	$g(L)$
−1.000	0.00	−0.50	−0.50	0.00
−0.866	0.07	−0.61	−0.62	−0.28
−0.707	0.15	−0.69	−0.74	−0.57
−0.500	0.25	−0.75	−0.87	−1.00
0.000	0.50	−0.83	−1.13	−2.53
+0.500	0.75	−0.82	−1.33	−6.69
+0.707	0.85	−0.80	−1.41	−12.40
+0.866	0.93	−0.78	−1.46	−28.80
+1.000	1.00	−0.75	−1.50	$-\infty$

13. Observations. For large crack lengths, say $L > 1$, (in units of the radius of the inner hole) the effect of the stress field caused by the circular hole is negligible insofar as the critical stress is concerned. In fact, the solution for a single crack in an infinite plate can be used provided that the single crack is compared with a crack of length $2(1 + L)$ for $k = 2$ and a length of $L + 2$ for $k = 1$.

On the other hand, for very small crack lengths the critical load appears to be governed primarily by the local stress field of the hole. This observation is made on the basis of the following argument: Since the tangential stress in the immediate vicinity of the hole where the crack is to be placed is $3T$ for the

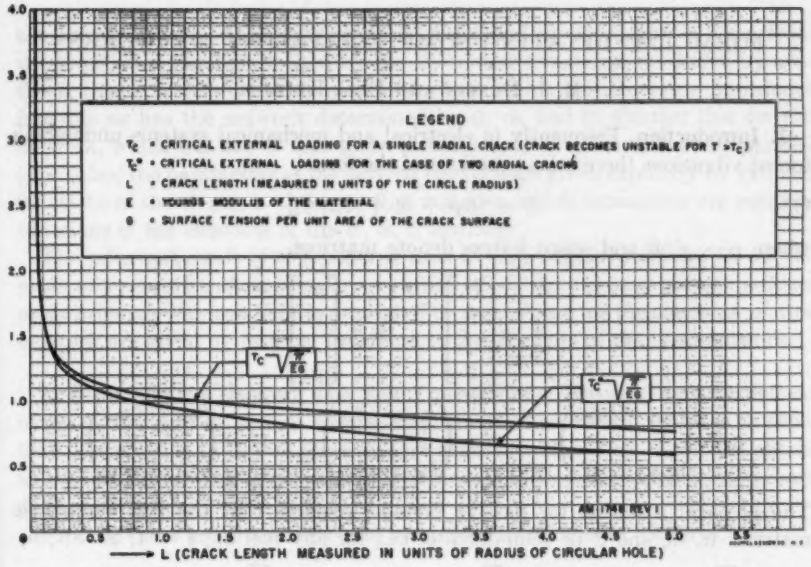


FIG. 3. Variation of the critical load with crack length for the case of simple tension at infinity.

case of simple tension and $2T$ for the case of all-around tension, one would expect that the ratio of the critical load for simple tension to the critical load for all-around tension would be two-thirds if governed by the local stress field. A comparison of the data of Figure 2 and Figure 3 indicates that such indeed is the case as the crack length approaches zero. (Note that the factor $(1 - \nu^2)$ must be neglected to make a consistent plane stress comparison.)

For crack lengths intermediate to those above, the behavior of the critical load does not appear capable of such simple physical arguments.

Finally, the critical internal pressure for the plane strain problem of a cylinder of very large wall thickness can be found simply by replacing T_c by P_c in Figure 2.

BIBLIOGRAPHY

[1] GRIFFITH, A. A., Phil. Trans. Royal Soc., London, A Vol. 221, p. 163 (1921)
[2] MUSHELISVILI, Mathematische Annalen, Vol. 107, pp. 282–312 (1932–33)
[3] TITCHMARSH, E. C., Theory of Functions, Oxford, p. 218 (1939)
[4] SOKOLNIKOFF, I. S., "Mathematical Theory of Elasticity", (Brown University Lecture Notes), pp. 243–319, (1941)
[5] MORKOVIN, V., Quart. Appl. Math., Vol. II, No. 4, pp. 350–352 (1944)
[6] OROWAN, E., Fatigue and Fracture of Metals, Wiley & Sons, Inc., New York, pp. 139–167, (1950)

WATERTOWN ARSENAL, WATERTOWN, MASS.

(Received August 17, 1954)

FORMULAE RELATING SOME EQUIVALENT NETWORKS*

R. J. Duffin and Elsa Keitzer

1. Introduction. Frequently in electrical and mechanical systems undergoing forced vibrations there arise equations of the form

$$\mathfrak{F} = (\mathfrak{A}p + \mathfrak{B} + \mathfrak{C}p^{-1})\mathfrak{Y} \tag{1}$$

where $p \sim d/dt$ and script letters denote matrices,

$$\mathfrak{F} = \begin{bmatrix} F_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \mathfrak{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$\mathfrak{A} = \begin{pmatrix} a_{11} \cdots a_{1n} \\ \vdots \\ a_{n1} \cdots a_{nn} \end{pmatrix}, \qquad \mathfrak{B} = \begin{pmatrix} b_{11} \cdots b_{1n} \\ \vdots \\ b_{n1} \cdots b_{nn} \end{pmatrix}, \qquad \mathfrak{C} = \begin{pmatrix} c_{11} \cdots c_{1n} \\ \vdots \\ c_{n1} \cdots c_{nn} \end{pmatrix}.$$

The physical situation for passive systems requires that the real symmetric matrices $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$ be semi-definite, i.e., for any real x_i, $i = 1, 2, \cdots, n$,

$$\sum_{i,j=1}^{n} a_{ij}x_ix_j \geqq 0, \qquad \sum_{i,j=1}^{n} b_{ij}x_ix_j \geqq 0, \qquad \sum_{i,j=1}^{n} c_{ij}x_ix_j \geqq 0.$$

(1) might be said to define an ideal network since it does not represent an actual electrical network without ideal transformers unless further restrictions on $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$ are met. $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$ may be considered as inductance, resistance, and elastance matrices respectively; $\mathfrak{F}$ and $\mathfrak{Y}$, as voltage and current matrices; and p, as $j \times$ (angular frequency).

The driving-point impedance of this ideal network may be defined as

$$Z = \frac{\text{determinant}(\mathfrak{A}p + \mathfrak{B} + \mathfrak{C}p^{-1})}{[\text{determinant}(\mathfrak{A}p + \mathfrak{B} + \mathfrak{C}p^{-1})]''} \tag{2}$$

where $[\text{determinant } (\mathfrak{A}p + \mathfrak{B} + \mathfrak{C}p^{-1})]''$ denotes the minor of the term in the first row, first column of the matrix $(\mathfrak{A}p + \mathfrak{B} + \mathfrak{C}p^{-1})$. It was shown by Bott and Duffin that there exists a network without transformers having this same driving-point impedance. In this paper we are interested in showing how this method may be carried out in algebraic form in simple cases.

We are concerned only with 2-loop networks, i.e., the case $n = 2$. Suppose that $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$ are real semi-definite matrices. First it is shown by means of a matrix transformation of the type employed by Howitt [1] that there is a network of a particular form which has the same impedance function Z as has (1). (Actually this step can be carried out for arbitrary n.) Let the inductance, resistance, and

* The work on this paper was sponsored by the Office of Ordnance Research, U. S. Army, Contract DA-36-061-ORD-378. The material in this paper is part of a thesis submitted by the second author to Carnegie Institute of Technology in partial fulfillment of the requirements for the Ph.D. degree.

elastance matrices of this new network be denoted by $\mathfrak{L}$, $\mathfrak{R}$, and $\mathfrak{S}$. Then from the impedance function defined by the $\mathfrak{L}$, $\mathfrak{R}$, and $\mathfrak{S}$ matrices we derive by synthesis theory a network with no transformers which has the same impedance function as has the network determined by $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$. Further this derived network, which has one of four given structures, is of the simple series-parallel type. Also, the parameters of the derived network are given explicitly by rational functions of the elements of the $\mathfrak{L}$, $\mathfrak{R}$, $\mathfrak{S}$ matrices, which themselves are rational functions of the elements of the $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ matrices.

The principal result of this paper is thus to show *explicit formulae for the elements of a network without transformers which has the same driving-point impedance as an arbitrary two-loop passive network*. The expressions for the elements of this network are given by rational functions of the elements of the ideal network.

2. Reduction to a Network with Real Transformers. The ideal network defined by $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ would in general require ideal transformers. First we want to show that it is possible by use of the transformation introduced by Howitt to obtain an equivalent network (as far as the impedance is concerned) which employs only real transformers.

Suppose there are given three real symmetric semi-definite matrices,

$$\mathfrak{A} = \begin{pmatrix} a_1 & a_2 \\ a_2 & a_3 \end{pmatrix}, \qquad \mathfrak{B} = \begin{pmatrix} b_1 & b_2 \\ b_2 & b_3 \end{pmatrix}, \qquad \mathfrak{C} = \begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}.$$

Then if $\mathfrak{M}$ is a matrix of the form

$$\mathfrak{M} = \begin{pmatrix} 1 & 0 \\ x & y \end{pmatrix}$$

with x and y real, and $\mathfrak{M}^*$ is its transpose

$$\mathfrak{M}^* = \begin{pmatrix} 1 & x \\ 0 & y \end{pmatrix},$$

the network with the inductance, resistance, and elastance matrices $\mathfrak{M}^*\mathfrak{A}\mathfrak{M}$, $\mathfrak{M}^*\mathfrak{B}\mathfrak{M}$, $\mathfrak{M}^*\mathfrak{C}\mathfrak{M}$, has the same impedance function as the network which $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ define [1].

First it is desired to determine x and y so that

$$\mathfrak{M}^*\mathfrak{A}\mathfrak{M} = \mathfrak{L} = \begin{pmatrix} l_1 & m \\ m & l_2 \end{pmatrix},$$

$$\mathfrak{M}^*\mathfrak{B}\mathfrak{M} = \mathfrak{R} = \begin{pmatrix} r_1 + r_2 & -r_2 \\ -r_2 & r_2 + r_3 \end{pmatrix},$$

$$\mathfrak{M}^*\mathfrak{C}\mathfrak{M} = \mathfrak{S} = \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix}.$$

Now

$$\mathfrak{M}^*\mathfrak{C}\mathfrak{M} = \begin{pmatrix} 1 & x \\ 0 & y \end{pmatrix}\begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ x & y \end{pmatrix}$$

$$= \begin{pmatrix} c_1 + 2c_2x + c_3x^2 & y(c_2 + c_3x) \\ y(c_2 + c_3x) & c_3y^2 \end{pmatrix}$$

$$= \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix},$$

$$\mathfrak{M}^*\mathfrak{B}\mathfrak{M} = \begin{pmatrix} b_1 + 2b_2x + b_3x^2 & y(b_2 + b_3x) \\ y(b_2 + b_3x) & b_3y^2 \end{pmatrix}.$$

$$= \begin{pmatrix} r_1 + r_2 & -r_2 \\ -r_2 & r_2 + r_3 \end{pmatrix}.$$

For $b_2c_3 - b_3c_2 \neq 0$, it is possible to select x and y so that $r_3 = 0$. This requires that

$$y[b_3y + (b_2 + b_3x)] = 0, \qquad y(c_2 + c_3x) = 0.$$

This gives

$$x = -c_2/c_3, \qquad y = [c_2b_3 - c_3b_2]/b_3c_2.$$

When $b_2c_3 - b_3c_2 = 0$, it isn't always possible to select x and y so that $r_3 = 0$, but in this case, x and y can be selected so that the synthesis is a simple applica-

TABLE 1

	Case 1 $b_3c_3 \neq b_3c_3$	Case 2 $b_2c_3 = b_3c_2$ $b_3, c_3 \neq 0$	Case 3 $c_3 \neq b_3 = b_3 = 0$	Case 4 $b_3 \neq c_3 = c_3 = 0$	Case 5 $c_3 = c_3 = b_3 = b_3 = 0$
x	$-\frac{c_2}{c_3}$	$-\frac{c_2}{c_3}$	$-\frac{c_2}{c_3}$	$-\frac{b_2 + b_3}{b_3}$	0
y	$\frac{b_2c_2 - b_2c_3}{b_3c_3}$	1	1	1	1
l_1	$a_1 + 2a_2x + a_3x^2$	$a_1 + 2a_2x + a_3x^2$	$a_1 + 2a_2x + a_3x^2$	$a_1 + 2a_2x + a_3x^2$	a_1
l_2	a_3y^2	a_3	a_3	a_3	a_3
m	$y(a_2 + a_3x)$	$a_2 + a_3x$	$a_2 + a_3x$	$a_2 + a_3x$	a_2
s_1	$\frac{c_1c_3 - c_2^2}{c_3}$	$\frac{c_1c_3 - c_2^2}{c_3}$	$\frac{c_1c_3 - c_2^2}{c_3}$	c_1	c_1
s_2	c_3y^2	c_3	c_3	0	0
r_1	$\frac{b_1b_3 - b_2^2}{b_3}$	$\frac{b_1b_3 - b_2^2}{b_3}$	b_1	$\frac{b_1b_3 - b_2^2}{b_3}$	b_1
r_2	b_3y^2	0	0	b_3	0
r_3	0	b_3	0	0	0

tion of the Brune theory. The case when $b_2c_3 - b_3c_2 \neq 0$ is more involved and will receive most of our attention. The values of x and y and the resulting elements of the $\mathfrak{L}$, $\mathfrak{R}$, $\mathfrak{S}$ matrices are shown in Table 1.

The network with $\mathfrak{L}$, $\mathfrak{R}$, $\mathfrak{S}$ as its inductance, resistance, and elastance matrices is shown in Figure 1.

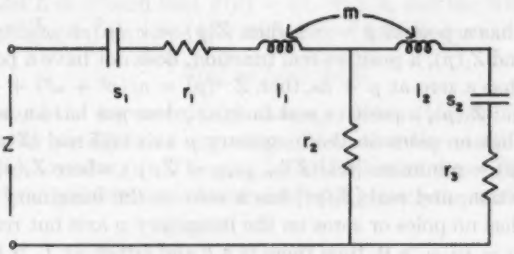


FIG. 1

3. Reduction to a Network Without Transformers. In the preceding section we have seen that we can always reduce the ideal network to the network shown in Figure 1. In this section we will carry out a synthesis to obtain an equivalent network without transformers. This will be at the expense of increasing the number of loops. Except for the case when $b_2c_3 - b_3c_2 = 0$, it is necessary to consider only the network shown in Figure 2. It will be seen that the exceptional case is easily dealt with.

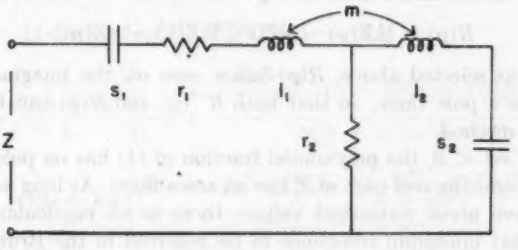


FIG. 2

The impedance function of the network of Figure 2 is

$$
\begin{aligned}
Z(p) &= [(l_1l_2 - m^2/l_2)p + r_1 + s_1p^{-1}] \\
&\quad + \frac{[r_2(l_2 + m)^2/l_2]p^2 + (s_2m^2/l_2)p + s_2r_2}{l_2p^2 + r_2p + s_2} \qquad (3) \\
&= [(l_1l_2 - m^2/l_2)p + r_1 + s_1p^{-1}] + Z'(p)
\end{aligned}
$$

It can be verified that $Z'(p)$ may be written in two other forms:

$$Z'(p) = [m(l_2 + m)/l_2]p - [(l_2 + m)p^2 + s_2][mp - r_2]/[l_2p^2 + r_2p + s_2], \qquad (4)$$

$$
\begin{aligned}
Z'(p) &= [-ms_2/l_2p] \\
&\quad + [(l_2 + m)p^2 + s_2][r_2(l_2 + m)p + s_2m]/l_2p(l_2p^2 + r_2p + s_2) \qquad (5)
\end{aligned}
$$

Using the synthesis procedures of Brune [2], Bott and Duffin [3], and another procedure which replaces one step in the Brune method, a network can be written down which is passive, has no transformers, and has the function $Z(p)$ given by (3) as its driving-point impedance. As long as $(l_2 + m) > 0$, the methods of Brune and Bott and Duffin are used to give the synthetic network. These methods are:

(a) If $Z(p)$ has a pole at $p = i\omega$, then $Z(p) = c/(p^2 + \omega^2) + Z_1(p)$, where $c > 0$ and $Z_1(p)$, a positive real function, does not have a pole at $p = i\omega$.

(b) If $Z(p)$ has a zero at $p = i\omega$, then $Z^{-1}(p) = c/(p^2 + \omega^2) + Z_1^{-1}(p)$, where $c > 0$ and $Z_1(p)$, a positive real function, does not have a zero at $p = i\omega$.

(c) If $Z(p)$ has no poles on the imaginary p axis and real $[Z(p)] > 0$ there, then $Z(p) = \text{minimum } [\text{real}(Z)]_{\text{im.axis}} + Z_1(p)$, where $Z_1(p)$ is a positive real function, and real $[Z_1(p)]$ has a zero on the imaginary p axis.

(d) If $Z(p)$ has no poles or zeros on the imaginary p axis but real $[Z(p)]$ has a zero at $p = i\omega$, $\omega > 0$, then there is a k and either an $L > 0$ or an $S > 0$ so that either

$$Z(i\omega) = i\omega L, \qquad Z(k) = kL,$$

or

$$Z(i\omega) = S/i\omega, \qquad Z(k) = S/k.$$

In either case,

$$Z(p) = [p/kZ(k) + 1/R(p)Z(k)]^{-1} + [R(p)/Z(k) + k/pZ(k)]^{-1},$$

where $R(p)$ is the Richard's function [4],

$$R(p) = [kZ(p) - pZ(k)]/[kZ(k) - pZ(p)].$$

With k as selected above, $R(p)$ has a zero on the imaginary p axis or $R(p)$ has a pole there, so that both $R^{-1}(p)$ and $R(p)$ can be reduced by Brune's method.

When $(l_2 + m) < 0$, the polynomial fraction in (4) has no poles on the imaginary p axis and the real part of Z has no zeros there. As long as the network parameters have given numerical values, there is no particular difficulty in determining that minimum resistance to be removed in the Brune process. In the case of this paper, however, the resistance may be one of several different functions of the network parameters, depending upon their interrelations. To avoid this difficulty, a different method is developed. Recall that a factor of $[(l_2 + m)p^2 + s_2]$ occurs in the polynomial fraction of (4). This term has the real roots $p = \pm[-s_2/(l_2 + m)]^{\frac{1}{2}}$. If k is selected as the positive real root, and L or S (whichever is >0) is selected as $Z(k)/k$ or $kZ(k)$ respectively, then these values of k and L or S can be used in the Richard's function for $Z_1(p)$ to give

$$Z_1(p) = [p/kZ_1(k) + 1/Z_1(k)R(p)]^{-1} + [R(p)/Z_1(k) + k/pZ_1(k)]^{-1},$$

where terms cancel in $R(p)$ to give a function which can be reduced by the Brune process.

It must be emphasized that this last procedure cannot in general replace the step of removing the minimum resistance. However, in certain cases it is valid.

Suppose that a rational impedance function has no poles on the imaginary p axis and its real part has no zeros there. Then this function has the form

$$Z(p) = f(p)/g(p)$$

where $f(p)$ and $g(p)$ have the same degree, say n. If n is an odd number, then there exist real k and L or S such that $Z(k) = kL$ or S/k, and the Richard's function formed using this positive value of k with L or S has the form

$$R(p) = f_1(p)/g_1(p)$$

where the sum of the degrees of $f_1(p)$ and $g_1(p)$ is less than $2n$. That this is true can be shown by an algebraic argument. (Note that the result here is a generalization of the Corollary to Theorem VI in Richard's paper [4].)

There are three possibilities to be considered in getting the synthetic network of Figure 2:

(i) $m > 0$; ($m = 0$ is trivial)

(ii) $-m > l_2$; ($-m = l_2$ is trivial)

(iii) $l_2 > -m > 0$.

In (i), $m > 0$, and (4) is used in the Bott-Duffin method with $L = m(l_2 + m)/l_2$, $\omega = [s_2/(l_2 + m)]^{\frac{1}{2}}$, $k = r_2/m$. We have

$$\begin{aligned}
kLR^{-1}(p) &= kL[kZ'(k) - pZ'(p)]/[kZ'(p) - pZ'(k)] \\
&= kL[k^2L - pZ'(p)]/[kZ'(p) - pkL] \\
&= L[-p - L(p^2 - k^2)/(Z'(p) - pL)] \\
&= \frac{m(l_2 + m)}{l_2}\left\{-p + \frac{\frac{m(l_2 + m)}{l_2}(p - k)(p + k)(l_2p^2 + r_2p + s_2)}{m(p - k)[(l_2 + m)p^2 + s_2]}\right\} \\
&= \frac{m(l_2 + m)}{l_2^2}\left\{\frac{(kl_2 + r_2)p^2 + \left(kr_2 + \frac{s_2m}{l_2 + m}\right)p + ks_2}{p^2 + s_2/(l_2 + m)}\right\} \\
&= \frac{m(l_2 + m)}{l_2^2}\left\{\frac{r_2(l_2 + m)}{m} + \frac{\left(\frac{r_2^2}{m} + \frac{s_2m}{l_2 + m}\right)p}{p^2 + s_2/(l_2 + m)}\right\} \\
&= R_2 + [p/S_2 + 1/L_2p]^{-1},
\end{aligned}$$

where

$$\begin{aligned}
R_2 &= r_2(l_2 + m)^2/l_2^2, \\
S_2 &= m(l_2 + m)[r_2^2/m + s_2m/(l_2 + m)]/l_2^2, \\
L_2 &= S_2(l_2 + m)/s_2.
\end{aligned}$$

$$\begin{aligned}
kLR(p) &= k^2L^2[R_2 + (p/S_2 + 1/L_2p)^{-1}]^{-1} \\
&= [1/R_1 + 1/(L_1p + S_1/p)]^{-1},
\end{aligned}$$

where

$$R_1R_2 = L_1S_2 = S_1L_2 = r_2^2(l_2 + m)^2/l_2^2 .$$

The network for Z, case (i), is shown in Figure 3. (The values of l_1, l_2, m, r_1, r_2, s_1, s_2 in terms of the a's, b's, c's are given in Table 1, Page 74.)

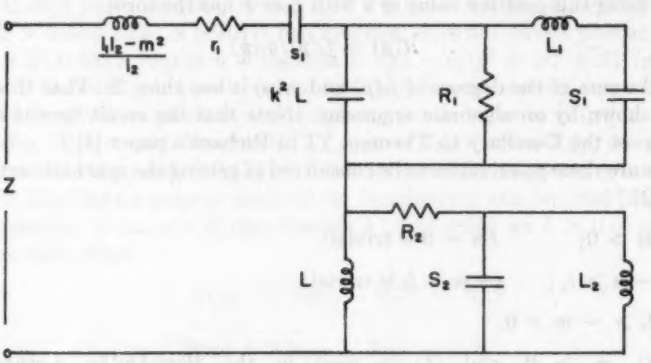


FIG. 3

In (ii), $-m > l_2$, and (4) is used with $L = m(l_2 + m)/l_2$, $k = [-s_2/(l_2 + m)]^{\frac{1}{2}}$ in the procedure outlined for $(l_2 + m) < 0$. We have

$$kLR^{-1}(p) = L[-p - L(p^2 - k^2)/(Z'(p) - pL)]$$

$$= \frac{m(l_2 + m)}{l_2}\left\{-p + \frac{\frac{m(l_2 + m)}{l_2}(p - k)(p + k)[l_2p^2 + r_2p + s_2]}{-(l_2 + m)(p - k)(p + k)[-mp + r_2]}\right\}$$

$$= \frac{m(l_2 + m)}{l_2^2}\left\{\frac{-r_2(l_2 + m)p - ms_2}{-mp + r_2}\right\}$$

$$= [1/R_3 + 1/L_2p]^{-1} + [1/R_4 + p/S_2]^{-1},$$

where

$$R_3 = r_2(l_2 + m)^2/l_2^2 = -r_2L_2/m,$$

$$S_2 = ms_2(l_2 + m)/l_2^2 = -r_2R_4/m.$$

$$kLR(p) = k^2L^2[(1/R_3 + 1/L_2p)^{-1} + (1/R_4 + p/S_2)^{-1}]^{-1}$$

$$= [(R_1 + S_1/p)^{-1} + (R_2 + L_1p)^{-1}]^{-1},$$

where

$$R_1R_2 = S_1L_2 = R_2R_4 = L_1S_2 = -mS_2 .$$

The network for Z, case (ii), is shown in Figure 4.

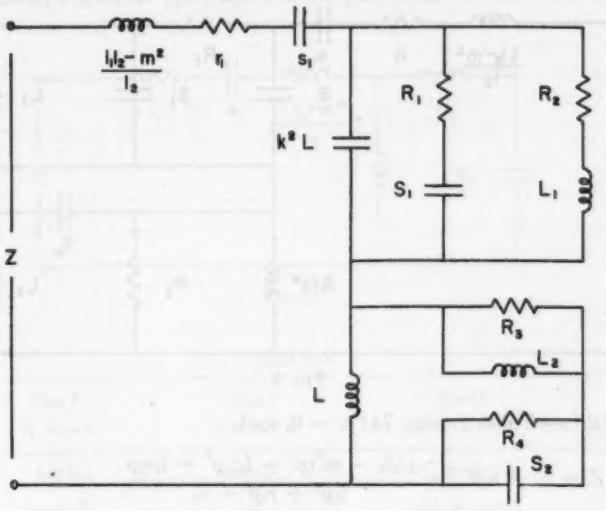


FIG. 4

In (iii), $l_2 > -m > 0$, and (5) is used in the Bott-Duffin method with $S = -ms_2/l_2$, $\omega = [s_2/(l_2 + m)]^{\frac{1}{2}}$, $k = -s_2m/r_2(l_2 + m)$. We have

$$\begin{aligned} R(p) &= [kZ'(p) - pZ'(k)]/[kZ'(k) - pZ'(p)] \\ &= -k/p + (p^2 - k^2)Z'(k)/p[Z'(p) - kZ'(k)/p]. \end{aligned}$$

$$\begin{aligned} SR(p)/k &= \frac{r_2(l_2 + m)}{l_2}\left\{-\frac{k}{p} + \frac{\frac{1}{p}(p - k)(p + k)\frac{S}{k}\, l_2p(l_2p^2 + r_2p + s_2)}{p[(l_2 + m)p^2 + s_2](p - k)r_2(l_2 + m)}\right\} \\ &= \frac{r_2(l_2 + m)}{l_2}\left\{\frac{l_2p^2 + (r_2 - km)p + s_2l_2/(l_2 + m)}{(l_2 + m)p^2 + s_2}\right\} \\ &= \frac{r_2(l_2 + m)}{l_2}\left\{\frac{l_2}{l_2 + m} + \frac{(r_2 - km)p}{(l_2 + m)p^2 + s_2}\right\} \\ &= R_1 + [p/S_1 + 1/L_1p]^{-1}, \end{aligned}$$

where

$$R_1 = r_2, \qquad S_1 = r_2(r_2 - km)/l_2, \qquad L_1 = S_1(l_2 + m)/s_2.$$

$$\begin{aligned} S/kR(p) &= [r_2^2(l_2 + m)^2/l_2^2]/[R_1 + (p/S_1 + 1/L_1p)^{-1}] \\ &= [1/R_2 + 1/(L_2p + S_2/p)]^{-1}, \end{aligned}$$

where

$$R_1R_2 = S_1L_2 = S_2L_1 = r_2^2(l_2 + m)^2/l_2^2.$$

The network for Z, case (iii), is shown in Figure 5.

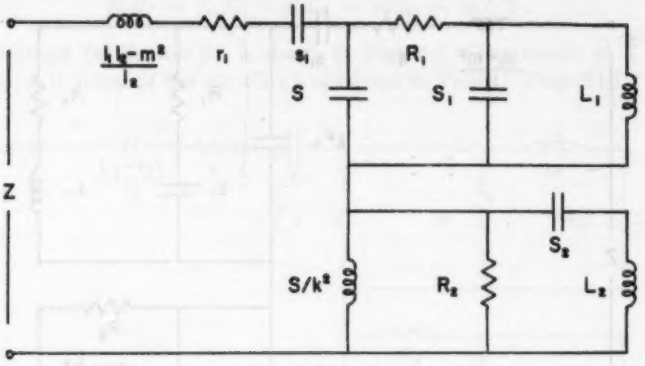


FIG. 5

In case 2 (see Table 1, page 74) $r_2 = 0$, and

$$Z = r_1 + s_1p^{-1} + \frac{(l_1l_2 - m^2)p^3 + l_1r_3p^2 + l_1s_2p}{l_2p^2 + r_3p + s_2}$$

$$= r_1 + s_1p^{-1} + \frac{l_1l_2 - m^2}{l_2}\,p + \frac{\frac{r_3m^2}{l_2}\,p^2 + \frac{s_2m^2}{l_2}\,p}{l_2p^2 + r_3p + s_2}$$

$$= r_1 + s_1p^{-1} + \frac{l_1l_2 - m^2}{l_2}\,p + \left\{\frac{l_2}{m^2p} + \frac{l_2^2p}{r_3m^2p + s_2m^2}\right\}^{-1}.$$

In case 3, $r_2 = r_3 = 0$, and

$$Z = r_1 + s_1p^{-1} + \frac{(l_1l_2 - m^2)p^2 + l_1s_2}{l_2p + s_2p^{-1}}$$

$$= r_1 + s_1p^{-1} + \frac{l_1l_2 - m^2}{l_2}\,p + \frac{\frac{s_2m^2}{l_2}}{l_2p + s_2p^{-1}}.$$

In case 4, $s_2 = r_3 = 0$, and

$$Z = r_1 + s_1p^{-1} + \frac{(l_1l_2 - m^2)p^2 + r_2(l_1 + l_2 + 2m)p}{l_2p + r_2}$$

$$= r_1 + s_1p^{-1} + \frac{l_1l_2 - m^2}{l_2}\,p + \frac{\frac{r_2}{l_2}\,(l_2 + m)^2p}{l_2p + r_2}.$$

In case 5, $r_2 = r_3 = s_2 = 0$, and

$$Z = r_1 + s_1p^{-1} + \frac{l_1l_2 - m^2}{l_2}\,p.$$

In cases 2–5, the synthetic networks are similar, so that all may be illustrated by one diagram, Figure 6. The values of the parameters in each case are given in Table 2.

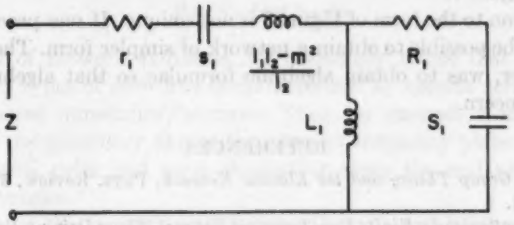


FIG. 6

TABLE 2

	Case 2 $b_2, c_2 \neq 0$	Case 3 $c_2 \neq b_2 = 0$	Case 4 $b_2 \neq c_2 = 0$	Case 5 $b_2 = c_2 = 0$
L_1	m^2/l_2	m^2/l_2	$(l_2 + m)^2/l_2$	0
R_1	r_2m^2/l_2^2	0	$r_2(l_2 + m)^2/l_2^2$	0
S_1	s_2m^2/l_2^2	s_2m^2/l_2^2	0	0

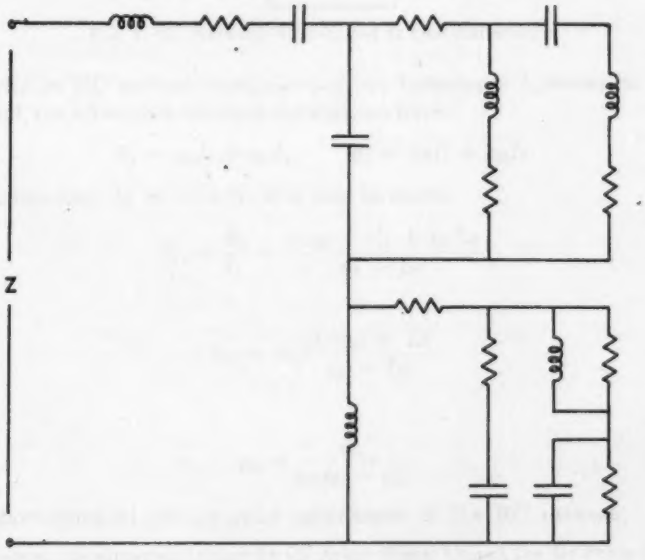


FIG. 7

Thus we see that the ideal network has an equivalent network without transformers with the same impedance. The equivalent network is of the form shown in Figure 3, 4, 5, or 6. It is clear that all of these networks can be represented in the form of Figure 7.

This reduction to the form of Figure 7 is not unique. If one proceeded numerically it might be possible to obtain a network of simpler form. The point of this paper, however, was to obtain algebraic formulae so that algebraic simplicity was also of concern.

REFERENCES

[1] N. Howitt, *Group Theory and the Electric Network*, Phys. Review, **37**, pp. 1583–1595 (1931).

[2] O. Brune, *Synthesis of a Finite Two-Terminal Network Whose Driving-Point Impedance is a Prescribed Function of Frequency*, J. Math. Phys., **10**, pp. 191–236 (1931).

[3] R. Bott and R. J. Duffin, *Impedance Synthesis Without Use of Transformers*, J. Applied Physics, **20**, p. 816 (1949).

[4] P. I. Richards, *A Special Class of Functions With Positive Real Part in a Half-Plane*, Duke Math. J., **14**, pp. 777–786 (1947).

Carnegie Institute of Technology

(Received March 8, 1954; Revised September 26, 1955)

AN EXISTENCE THEOREM FOR DRIVING-POINT IMPEDANCE FUNCTIONS*

By N. DeClaris

In the field of network synthesis, we frequently notice that the structural form of certain types of networks is characterized by specific zero-pole patterns of their associated impedance functions. Thus, for example, a simple zero-pole alternance on the imaginary axis of the complex-frequency plane implies an LC network, whereas poles and zeros alternating along the real axis characterize RC or RL networks.

The purpose of this paper is to determine the pole-zero pattern for a driving-point impedance realizable as an RC network terminated to an inductance. In essence, this situation is similar to that arising in the well-known method of Darlington (1), who showed that any realizable RLC driving-point impedance function can be synthesized as an LC network terminated to a resistance. The results obtained are, however, not quite so general as Darlington's since not all RLC driving-point impedance functions can be realized in this manner.

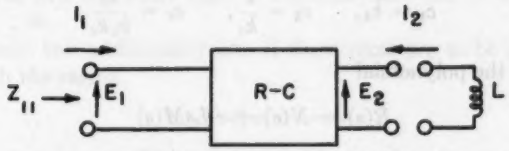


Fig. 1. RC Network Terminated to One Inductance

Consider an RC network terminated in an inductance L, shown in Fig. 1. In general, for a two-pair terminal network we have

$$E_1 = z_{11}I_1 + z_{12}I_2\,, \qquad E_2 = z_{12}I_1 + z_{22}I_2 \tag{1}$$

Since, in this case, $E_2 = -Ls\,I_2$, it is easy to derive

$$Z_{11} = \frac{E_1}{I_1} = \frac{z_{11}z_{22} - z_{12}^2 + z_{11}Ls}{z_{22} + Ls} \tag{2}$$

or

$$Z_{11} = z_{11}\,\frac{(1/y_{22}) + Ls}{z_{22} + Ls} \tag{3}$$

where

$$y_{22} = \frac{z_{11}}{z_{11}z_{22} - z_{12}^2} \tag{3a}$$

is the short-circuited driving-point admittance of the RC network.

* This work was supported in part by the Army (Signal Corps), the Air Force (Office of Scientific Research, Air Research and Development Command), and the Navy (Office of Naval Research).

It is well known that for an RC network the driving-point impedance and admittance have simple real negative poles and zeros which alternate along the real axis of the s-plane. Therefore

$$z_{11} = \frac{P(s)}{Q(s)} k_1 = k_1 \frac{\Pi_{i=1}^{n'}(s + p_i)}{\Pi_{i=1}^{m'}(s + q_i)} \tag{4}$$

where $n' = m'$, or $n' + 1 = m'$, and $q_1 < p_1$. Similarly,

$$z_{22} = k_2 \frac{D(s)}{Q(s)} = k_2 \frac{(s + d_1)(s + d_2)\cdots}{(s + q_1)(s + q_2)\cdots} \quad q_1 < d_1 \tag{5}$$

$$y_{22} = k_3 \frac{M(s)}{N(s)} = k_3 \frac{(s + m_1)(s + m_2)\cdots(s + m_m)}{(s + n_1)(s + n_2)\cdots(s + n_n)} \quad m_1 < n_1 \tag{6}$$

By substitution, Eq. 3 becomes

$$Z_{11} = \frac{P(s)}{M(s)} \frac{N(s) + c_1 sLM(s)}{D(s) + c_2 sLQ(s)} c_3 \tag{7}$$

where

$$c_1 = k_3, \qquad c_2 = \frac{1}{k_2}, \qquad c_3 = \frac{k_1}{k_2 k_3}$$

Let us study the polynomial

$$R(s) = N(s) + c_1 LsM(s) \tag{8}$$

$N(s)$ and $M(s)$ have all of their roots real negative ($-n_1, -n_2, \cdots, -n_n$, and $-m_1, -m_2, \cdots, -m_m$) which alternate when arranged in a sequence according to their ascending magnitude. A sketch of $N(s)$ and $M(s)$ is given in Fig. 2 for s real and negative. Both $M(s)$ and $N(s)$ start positive, since $M(0) = m_1 m_2 m_3 \cdots m_m$, and $N(0) = n_1 n_2 n_3 \cdots n_n$.

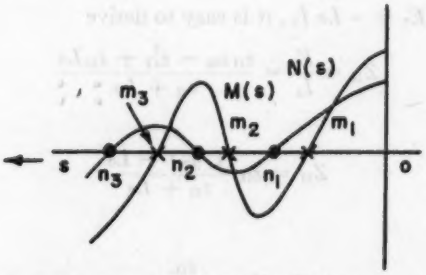


FIG. 2. The Polynomials $M(s)$ and $N(s)$ for real s

A plot of the polynomial $s\,M(s)$ is given in Fig. 3. Observe that $s\,M(s) < 0$ for $0 > s > -m_1$, since $M(s)$ is positive for that region and s is negative real.

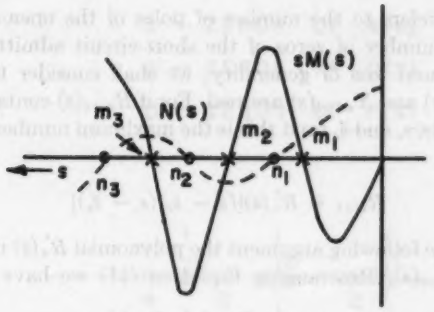


FIG. 3. The Polynomial $sM(s)$ for real s

According to the fundamental theorem of algebra (2), the total number of roots of the polynomial $R(s)$ is $m + 1$. By inspection of Fig. 3, it is obvious that regardless of the magnitude of N, M, and L there will be one, and only one, zero of $R(s)$ in the region $-n_i > s > -m_{i+1}$, where $i = 1, 2, \cdots, m$. Therefore, regardless of the value of L the polynomial $R(s)$ will always have $m - 1$ real and negative roots. No real roots exist in the intervals $-m_i > s > -n_i$, where $i = 1, 2, \cdots, m$.

There remain two additional roots. If these roots are to be real, at least one must occur in the region

$$0 < s < -m_1 \tag{9}$$

However, by reference to Fig. 3 it is possible to select a value L such that no real root of $R(s)$ exists in the interval $0 < s < -m_1$. Hence, by elimination, the two remaining roots must be complex conjugates. A similar analysis can be carried out for the polynomial

$$X(s) = D(s) + c_2 L s Q(s) \tag{10}$$

As a result the following can be stated.

THEOREM. The driving-point impedance of an RC network terminated to an inductance will have at the most one pair of complex conjugate zeros and one pair of complex conjugate poles.

COROLLARY 1. The driving-point impedance of an RL network terminated to a capacitor will have at the most one pair of complex conjugate zeros and one pair of complex conjugate poles.

The proof is identical to that in the case of an RC network terminated to an inductance.

Thus far it has been shown that the polynomials $R(s)$ and $X(s)$ at the most will have a pair of complex roots each. We shall establish the set order of the remaining real roots. Consider the equation

$$Z_{11}(s) = \frac{P_{m'}(s)}{M_m(s)} \frac{R_{m+1}(s)}{X_{m'+1}(s)} \tag{11}$$

The subscript m' refers to the number of poles of the open-circuit impedance $z_{11}(s)$, and m the number of zeros of the short-circuit admittance $y_{22}(s)$ of the RC network. Without loss of generality, we shall consider the case where all the zeros of $R_{m+1}(s)$ and $X_{m'+1}(s)$ are real. For if $R_{m+1}(s)$ contains a pair of complex conjugate roots, s_c and $\bar{s}_c$ (and this is the maximum number it can ever have) one simply writes:

$$R_{m+1} = R'_n(s)[(s - s_c)(s - \bar{s}_c)] \tag{12}$$

and considers in the following argument the polynomial $R'_n(s)$ in place of $R_{m+1}(s)$. Similarly for $X_{m'+x}(s)$. Rearranging Equation (11) we have

$$Z_{11}(s) = \frac{R_{m+1}(s)}{M_m(s)} \frac{P_{m'}(s)}{X_{m'+1}(s)} \tag{13}$$

Let $-r_i$, $-m_i$, $-p_i$, $-x_i$, $(i = 1, 2, 3, \cdots)$ be the roots of the polynomials $R_{m+1}(s)$, $M_m(s)$, $P_{m'}(s)$, and $X_{m'+1}(s)$ respectively. It has been shown, Fig. 3, that if one arranges m_i and r_i according to their ascending magnitude the following sequence is formed

$$r_1, m_1, r_2, m_2, \cdots, m_m, r_{m+1} \tag{14}$$

If we are considering $R'_n(s)$ in place of the polynomial $R_{m+1}(s)$, then r_1 and r_{m+1} are missing. Similarly the sequence*

$$x_1, q_1, x_2, \cdots, q_{m'}, x_{m'+1} \tag{15}$$

follows from Equation (10) where $-q_i$ are the roots of the polynomial $Q(s)$, and from Equation (4)

$$q_1, p_1, q_2, \cdots, q_{m'}, p_{m'} \tag{16}$$

We shall now show that x_i and p_i form a sequence in which no more than two x_i or two p_i can occur in succession. From the sequence of (15) it follows that between q_k and q_{k+1} there is one and only one x_k and from the sequence (16) it follows that in the same interval q_k and q_{k+1} there is one and only one p_k. Now x_k and p_k can be in one of the two orders $x_k p_k$ or $p_k x_k$. From the same argument, x_{k+1} and p_{k+1} will occur between the bounds q_{k+1} and q_{k+2}, and they will be in the order $x_{k+1}p_{k+1}$ or $p_{k+1}x_{k+1}$. It is clear, therefore, that we can have the four possible subsequences

$$x_k\ q_k\ x_{k+1}\ q_{k+1}\ ; \qquad q_k\ x_k\ x_{k+1}\ q_{k+1}$$

$$x_k\ q_k\ q_{k+1}\ x_{k+1}\ ; \qquad x_k\ q_k\ x_{k+1}\ q_{k+1}$$

in the sequence of x_i and p_i. However, at any rate no more than two x_i or p_i can occur in succession. The same result can most easily be seen from Fig. 4.

* Herewith by the term "sequence" we shall mean the arrangement of two sets of real numbers in order of their ascending magnitude.

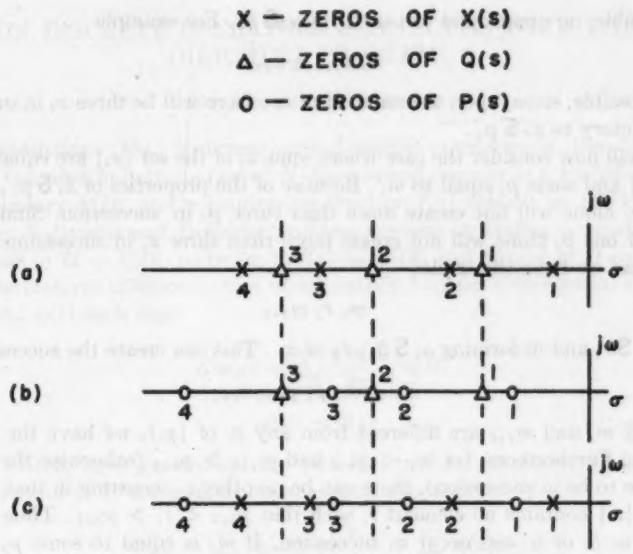


FIG. 4. Relation of the zeros of $X(s)$, $Q(s)$, $P(s)$

Fig . 4c shows only a part of the pole-zero sequence of the driving-point impedance function $z_{11}(s)$. To complete the pole-zero pattern, one has to insert, see Equation (13), in their appropriate order, the poles and zeros of the rational function $R_{m+1}(s)/M_m(s)$. We shall now introduce a new notation. Let

$x_i \mathbf{S} p_i$ denote the sequence of the sets $\{x_i\}$ and $\{p_i\}$

$m_i \mathbf{S} r_i$ denote the sequence of the sets $\{m_i\}$ and $\{r_i\}$

$\alpha_i \mathbf{S} \beta_i$ denote the sequence of the sets $\{\alpha_i\}$ and $\{\beta_i\}$

where $\{\alpha_i\} = \{x_i\} \mathbf{U} \{m_i\}$ and $\{\beta_i\} = \{p_i\} \mathbf{U} \{r_i\}$.

Clearly from Equation (13)

$$\alpha_i \mathbf{S} \beta_i = x_i \mathbf{S} p_i \mathbf{U} m_i \mathbf{S} r_i \tag{17}$$

Thus far we have established the following:

(a) $x_i \mathbf{S} p_i$ consists of succession of pairs $x_i p_i$ and $p_i x_i$ with the possibility of termination in either a single x_i or a single p_i

(b) $m_i \mathbf{S} r_i$ consists of succession of pairs either $m_i r_i$ or $r_i m_i$

For x_i different from r_i and p_i different from m_i, it follows from Equation (17) that no more than three α_i or β_i can occur in succession. That is, where the triplets

$$x_j\, x_{j+1}\, r_k \qquad x_j\, r_k\, x_{j+1}$$

are possible, no quadruplet can occur in $\alpha_i \mathbf{S} \beta_i$. For example

$$x_j\, x_{j+1}\, r_k\, x_{j+2}$$

is not possible, since, upon the removal of r_k, there will be three x_i in succession, contradictory to $x_i \mathbf{S} p_i$.

We shall now consider the case where some x_i of the set $\{x_i\}$ are equal to some r_i of $\{r_i\}$ and some p_i equal to m_i. Because of the properties of $x_i \mathbf{S} p_i$, removal of one x_i alone will not create more than three p_i in succession. Similarly, removal of one p_i alone will not create more than three x_i in succession.

Suppose we have the triplet

$$m_j\, r_j\, m_{j+1}$$

from $m_i \mathbf{S} r_i$ and in forming $\alpha_i \mathbf{S} \beta_i$, $r_j = x_k$. This can create the succession

$$x_{k-1}\, p_{k-1}\, p_k\, p_{k+1}\, x_{k+1}$$

Now if m_j and m_{j+1} are different from any p_i of $\{p_i\}$, we have the previous situation. Furthermore, for $m_j < p_{k-1}$ and $m_{j+1} > p_{k+1}$ (otherwise the three p_i will cease to be in succession), there can be no other r_i occurring in that interval. That is $\{r_i\}$ contains no element r_i such that $p_{k-1} < r_i > p_{k+1}$. Thus no more than three β_i or α_i can occur in succession. If m_j is equal to some p_p of $\{p_i\}$, then the situation repeats itself and can create no more than three α_i or three β_i in succession until we reach distinct and different element values of $x_i \mathbf{S} p_i$ and $m_i \mathbf{S} r_i$, a situation that we have previously considered. Thus in all cases no more than three α_i or three β_i can occur in succession in $\alpha_i \mathbf{S} \beta_i$. We have therefore proved

COROLLARY 2. The real poles and zeros of an RC (or RL) driving-point impedance terminated to an inductance (or capacitance) form a set when arranged according to their ascending magnitude in which no more than three poles or three zeros can occur in succession.

Acknowledgment. The writer wishes to express his appreciation to Professor E. A. Guillemin for his suggestions and criticisms during the preparation of this paper.

It is also a pleasure to acknowledge the assistance and encouragement I received from Professor R. E. Scott.

REFERENCES

1. S. J. DARLINGTON, J. Math. Phys. **28**, 280 (1939).
2. J. L. WALSH, Location of the critical points of rational functions in the complex domain, Am. Math. Soc. Colloquium Publication, **34** (1950).

DEPARTMENT OF ELECTRICAL ENGINEERING
AND RESEARCH LABORATORY OF ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

(Received March 8, 1955)

ON DISCRETE OPERATORS CONNECTED WITH THE DIRICHLET PROBLEM[1, 2]

By Mervin E. Muller

1. Introduction. The N-dimensional Dirichlet problem is the following. Given a bounded finitely connected N-dimensional domain in a Euclidean space with boundary $\Gamma[D]$, and a continuous function $f(x)$ defined on the boundary $\Gamma[D]$, the N-dimensional Dirichlet problem consists in finding a function $u(x)$ continuous in $D + \Gamma[D]$, reducing to $f(x)$ on $\Gamma[D]$, and having in D continuous partial derivatives of second order which satisfy Laplace's differential equation, i.e., to find $u(x)$ such that

$$\Delta^2 u(x) = \sum_{i=1}^{N} \frac{\partial^2 u(x)}{\partial x_i^2} = 0 \qquad x \in D \tag{1.1}$$

$$u(x) = f(x) \qquad x \in \Gamma[D]. \tag{1.2}$$

Some writers have generalized the Dirichlet problem to allow the following: i.e., in place of (1.1) the general linear elliptic operator

$$L[u] = \sum_{i,j=1}^{N} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \, \partial x_j} + \sum_{j=1}^{N} b_j(x),$$

is introduced where $a_{ij}(x)$, $b_j(x)$ possess continuous second order derivatives in D, while in place of (1.1) they have $u(x) = g(x)$, $x \in \Gamma[D]$, where $g(x)$ is permitted to have points of discontinuity on $\Gamma[D]$. These generalizations of the Dirichlet problem will not be considered in this paper even though some of the material developed is applicable to the more general problem. Since the one-dimensional Dirichlet problem reduces to finding a straight line through two points it will be completely omitted from subsequent consideration in this paper.

The contribution that any numerical study might make to boundary value problems when the analytical questions of existence and uniqueness of a solution have not been settled is a moot question. Consequently, we make the following assumption:

Fundamental Assumption. Throughout this paper the boundary $\Gamma[D]$ will be assumed to be of sufficient regularity to ensure that the Dirichlet problem has a solution and that it is unique. Kellogg [17] and [18] has a detailed discussion and bibliography concerning regularity conditions on $\Gamma[D]$; see also de La Vallée Poussin [33]. The term "Monte Carlo" is being used here as it has been used by others, e.g. Metropolis and Ulam [23], in order to convey that

[1] Sections of this paper are part of a Doctoral Dissertation written at the University of California, Los Angeles. The research was sponsored in part by the National Bureau of Standards while the author was a Fellow in Mathematics at the Institute for Numerical Analysis.

[2] The preparation of this paper was sponsored (in part) by the Office of Naval Research.

an unknown solution to a given physical problem is being estimated by a method which essentially depends on a statistical sampling technique.

2. Operator and Network Considerations. 2.1. *Introduction.* The classical approach of Numerical Analysis has been to attempt to solve a given boundary value problem numerically by the method of "finite differences," e.g., see Milne [24]. This involves replacing the given domain and boundary by a predetermined network of points and then replacing the differential or integral operator by an appropriate difference operator. In contemplating a numerical technique using finite differences it seems natural to ask what would be the "best" procedure of this type to employ. The purpose of this section is to consider this question. It has been customary to select a specific type of network of points. Then, relative to this network of points, an approximation to the operator in question is derived. Here we will reverse the order of procedure by studying first how "best" to approximate the given differential or integral operator without reference to a specific predetermined network of points. This seems to be a reasonable approach when considering a problem using a high speed digital computer since both the difference operator and network need to be programmed for the particular computer.

Though the application of this procedure will be restricted to the Dirishlet problem the germ of the idea can be used for other finite difference problems. Essentially it will be necessary to expand the operator under investigation into a series of "difference operators" with arbitrary coefficients. Then with the desired order of approximation specified, the selection of the associated finite difference operator subject to a proper set of points is made. This collection of points will be referred to as a "neighbor set." One of our purposes will be to search for "best" neighbor sets. For application we would then construct, whenever possible, a network of points consisting of "best neighbor sets."

This approach is suggested in order to facilitate the formulation of mathematical criteria which will serve as a rational basis for selecting a "best" network of points. The approach seems promising for higher dimensional problems which will depend for their solution on the use of computing equipment. For the two dimensional Dirichlet problem it is seen in section 2.3 that this approach adds further justification for the networks already suggested, see for example Collatz [5], Milne [24], and Southwell [29]. Grinter [14] contains a very extensive bibliography of such numerical methods.

Other authors sometimes refer to a network of points as a grid or lattice of points. However we will use the term lattice only in those situations where the network of points actually satisfies the following technical definition of a lattice.

DEFINITION 2.1. *N-dimensional lattice.* An N-dimensional lattice consists of all points $P = \sum_{i=1}^{N} \lambda_i e_i$, where $\lambda_1, \cdots, \lambda_N$ are integers and $e_1, \cdots, e_N$ are given independent vectors in an N-dimensional Euclidean space.

DEFINITION 2.2. *Neighbors on a lattice.* Two points P_1 and P_2 are neighbors on a lattice if $\| P_1 - P_2 \| = e_i$ for some i, where $P_j = \sum_{i=1}^{N} \lambda_i^{(j)} e_i$, $j = 1, 2$.

Since we will study approximations to an operator in a neighborhood of a given point it is convenient to refer to the discrete collection of points used in the approximation as the "neighbor set" of the point. For the sake of emphasis we will now give a formal definition of a neighbor set.

DEFINITION 2.3. *N-dimensional neighbor set.* If there is a rule for associating a specific set of points in an N-dimensional space with an arbitrary point x_0 of the space then this set of points will be referred to as the neighbor set of x_0. If the points of the neighbor set form an N-space then the set is referred to as an N-dimensional neighbor set of x_0.

The rules for specifying a particular type of neighbor set will be made in subsequent pages. It will be seen that in a Monte Carlo solution of a boundary value problem a neighbor set of a given point represents the possible positions from the given point after a single transition in the appropriate Markov chain. In a Relaxation technique a neighbor set of a given point would denote the set of points relative to the given point about which a single step of the relaxation process would take place. Concerning Relaxation techniques, see for example Southwell [29].

2.2. *Preliminary definitions for the discrete Dirichlet problem.* When replacing the N-dimensional Dirichlet problem by a discrete approximation it has been conventional to seek discrete approximations to Laplace's operator on a given specified network of points, e.g., Collatz [5], Milne [24], Phillips and Wiener [27], and Southwell [29]. Bickley [3] and Albrecht [1] together constitute a study of how to improve the order of the approximation to the Laplacian in the two dimensional case when using certain fixed networks.

However in studying approximations to the operator *without* reference to a predetermined network of points we will need to consider criteria from which it will be possible to decide on an appropriate approximation. The task of finding criteria which will serve as a basis for selecting an appropriate discrete approximation for the Dirichlet problem will be somewhat simplified here if we make use of the fact that the desired solution is harmonic in its domain D.

Since the solution to the Dirichlet problem is a harmonic function in the domain D the function has the surface and volume mean value properties, see for example Courant and Hilbert [8], page 249. Thus the value of the harmonic function at the point x_0 which belongs to D is equal to the mean of its values, either on the surface or over the volume, of any N-dimensional sphere contained in D with center at x_0. Hence rather than determining discrete analogs to Laplace's operator directly, it is advantageous first to consider discrete analogs to the integral mean.

The latitude present when formulating a discrete mean value approximation is considerable. Before delineating the class of discrete mean value approximations so that they will possess or reflect certain desirable attributes, we consider the following general definitions.

DEFINITION 2.4. *A general discrete mean value.* The operator $\bar{U}_k(x_0, r, v, w)$ is a k-point mean value approximation about the point x_0 to the integral mean about x_0 for all functions u that are harmonic in the domain D of the N-dimen-

sional Dirichlet problem if

$$\bar{U}_k(x_0, r, v, w) = \sum_{j=1}^{k} w_j u(x_j), \qquad k \geqq N+1, \tag{2.1}$$

where the w_j's are considered as weights satisfying the following

$$w_j \text{ is a real number for } j = 1, 2, \cdots, k \text{ and } \sum_{j=1}^{k} w_j = 1, \tag{2.2}$$

where the k points constituting the neighbor set of x_0 belong to D and

$$x_j = x_0 + r_j v_j, \qquad j = 1, 2, \cdots, k, \tag{2.3}$$

and where the r_j's are scalars and the v_j's denote fixed unit vectors in N-space with coordinates v_{jl}, $l = 1, 2, \cdots, N$.

DEFINITION 2.5. *k-point harmonic approximation of m^{th} order.* Subject to the conditions of definition 2.4 the operator $\bar{U}_k(x_0, r, v, w)$ will be called a k-point harmonic approximation of order m if $\bar{U}_k(x_0, r, v, w) = u(x_0) + 0(\hat{r}^m)$ for *all* functions u harmonic in D and where $\hat{r} = \min\{r_j, j = 1, 2, \cdots, k\}$. $0(\hat{r}^m)$ signifies that as $\hat{r} \to 0$, $|0(\hat{r}^m)| < C\hat{r}^m$, where C is a positive constant.

The approach will be to study the discrete operator $\bar{U}_k(x_0, r, v, w)$ at a given point x_0. For a given value of k it is not immediately clear what arrangement of the k points constituting the neighbor set would be "best." Nor is it immediately clear what values of the weights would be "best."

Intuitively it is reasonable to feel that: (1) the approximation should in the limit possess the properties possessed by the continuous problem, (2) the neighbor set should be as compact as possible to avoid increasing the numerical complications at the boundary of the given domain, and (3) the neighbor set should consist of as few points as possible to save machine time, if possible, but enough points to insure a reasonable order of approximation.

Thus we formulate the following criteria.

DEFINITION 2.6. *Optimum harmonic approximation.* A k-point harmonic approximation with a given value of $\hat{r}$ as specified by definition 2.5 will be *optimum* at x_0 if no other harmonic approximation at x_0, with the same value of $\hat{r}$, based on k or less points has as high an order of approximation.

DEFINITION 2.7. *Uniformly optimum harmonic approximation.* A k-point harmonic approximation will be a uniformly optimum harmonic approximation over a set of points if it is an optimum approximation at each point of the set and if the order of the approximation is the same for each point of the set.

We will study the question of optimum harmonic approximations for a fixed value of k in two situations. In the first case we will specify that the k points are all at the same distance from x_0, i.e., $r_j = \hat{r}$, $j = 1, 2, \cdots, k$. We then seek how the k points should be oriented, i.e., determine the vectors v_j, $j = 1, 2, \cdots, k$, and also seek how the weights w_j, $j = 1, 2, \cdots, k$, should be selected. In the second case we will specify that all the weights are uniform, i.e., $w_j = 1/k$, $j = 1, 2, \cdots, k$. We then seek how the r_j's and v_j's, $j = 1, 2, \cdots, k$ should be selected.

If a network of points is formed from neighbor sets we will observe the following conventions. If the network of points is a lattice then an *interior point*

of this network follows the customary requirement of having all its neighbors included in the domain D. If the network is not a lattice then an *interior point* of this network must have its neighbor set included in the domain D.

In addition to the interior and boundary points of the network there may be a set of points which we will call intermediate points.

DEFINITION 2.8. *Intermediate points.* An intermediate point of a given network is any point of the network which is contained in the given domain D but has the addition property that one or more members of its neighbor set relative to the network are not in D.

Clearly, unless $\Gamma[D]$ is of a very special nature intermediate points will occur on the network. It is hoped that the use of the nomenclature intermediate point will add clarity in the subsequent discussions. Since an intermediate point is not an interior point some may prefer to look upon these points as boundary points. However it is precisely at the so-called intermediate points where one finds an interpolation scheme used, or at least a subsequent change in the behavior of the random walk or relaxation method.

We next consider the following definition.

DEFINITION 2.9. *Admissible network.* A network of points will be called admissible for a given problem if there exists a uniformly optimum approximation method on the interior points of this network.

In passing it is interesting to note that in the event $u(x_0) = \overline{U}_k(x_0, r, v, w)$, with $r_j = r$ and $w_j = 1/k$, $j = 1, 2, \cdots, k$, at each point of the network then in the two dimensional case the function is a harmonic polynomial and the works of Walsh [34] and Beckenbach and Reade [2] are applicable. When $k = 2N$ and the interior points form a lattice whose basic cell is an N-dimensional hyper-cube and when $u(x_0) = \overline{U}_k(x_0, r, v, w)$, with $r_j = r$ and $w_j = 1/2N$, $j = 1, 2, \cdots, 2N$, then on these points $u(x_0)$ is called a discrete harmonic function by Heilbroon [15] and he has studied some of the properties of u.

The definitions of this section were formulated in a manner such that the questions of optimum approximations and admissible networks could be translated to questions concerning "best" neighbor sets. Each type of boundary value problem it is hoped will have its classes of "best" or admissible networks depending on the order of approximation desired.

2.3. *The two dimensional discrete Dirichlet problem.* The two dimensional case is treated separately as it is not a special case of the N-dimensional problem in that the techniques employed and the results obtained do not all generalize for higher dimensions.

THEOREM 2.1. *In a two dimensional problem a harmonic approximation $\overline{U}_k(x_0, r, v, w)$ which uses k points all at the same distance r from x_0 is optimum if and only if the weights $w_j = 1/k$ for $j = 1, 2, \cdots, k$ and the k points constituting the neighbor set are restricted to be the vertices of a regular inscribed polygon of k sides with center x_0 and radius r. The optimum order of approximation is k.*

Proof. By definition of $\overline{U}_k(x_0, r, v, w)$, (definition 2.4), and by the conditions of this theorem the k points are chosen such that $x_j \in D$ and $\| x_j - x_0 \| = r$, $j = 1, 2, \cdots, k$. Since the class of functions being considered is harmonic it

will be possible to have a Taylor series representation about x_0 for each of the k values of any arbitrarily selected harmonic function u, i.e.,

$$u(x_j) = u(x_0) + \sum_{m=1}^{\infty} c_m r_j^m e^{im\theta_j}. \tag{2.4}$$

Since $\overline{U}_k(x_0, r, v, w) = \sum_{j=1}^{k} w_j u(x_j)$, $\sum_{j=1}^{k} w_j = 1$, and $r_j = r, j = 1, 2, \cdots, k$, we have

$$\overline{U}_k(x_0, r, v, w) = u(x_0) + \sum_{m=1}^{\infty} c_m \left(\sum_{j=1}^{k} w_j r^m e^{im\theta_j}\right). \tag{2.5}$$

It will now be shown that by a proper choice of the w_j's and θ_j's, $j = 1, 2, \cdots, k$, that the terms involving r^m, $m = 1, 2, \cdots, k-1$, can be made to vanish simultaneously, i.e., we have

$$\sum_{j=1}^{k} w_j e^{im\theta_j} = 0, \qquad \text{for } m = 1, 2, \cdots, k-1. \tag{2.6}$$

Since $\hat{r} = r$, (see definition 2.5), it is clear then if (2.6) holds for $m = 1, 2, \cdots, k-1$ that the k-point harmonic approximation is optimum in the sense of definition 2.6. Further the optimum order of approximation is k.

Consider the following vectors q_m, $m = 1, 2, \cdots, k-1$, where the coordinates of q_m are $\sqrt{w_j}\, e^{im\theta_j}$, $j = 1, 2, \cdots, k$, and the vector q_k has coordinates $\sqrt{w_j}$, $j = 1, 2, \cdots, k$. We will assume without any loss in generality that $w_j > 0, j = 1, 2, \cdots, k$. Only minor modifications are needed if we allow any of the w_j's to take on negative values. However we would ultimately arrive at the same results, namely, that $w_j = 1/k, j = 1, 2, \cdots, k$.

Now consider the following matrix Q of row vectors.

$$Q = \begin{bmatrix} \sqrt{w_1}e^{i\theta_1} & \sqrt{w_2}e^{i\theta_2} & \cdots & \sqrt{w_k}e^{i\theta_k} \\ \sqrt{w_1}e^{i2\theta_1} & \sqrt{w_2}e^{i2\theta_2} & \cdots & \sqrt{w_k}e^{i2\theta_k} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sqrt{w_1}e^{i(k-1)\theta_1} & \sqrt{w_2}e^{i(k-1)\theta_2} & \cdots & \sqrt{w_k}e^{i(k-1)\theta_k} \\ \sqrt{w_1} & \sqrt{w_2} & \cdots & \sqrt{w_k} \end{bmatrix}.$$

Since $\sum_{j=1}^{k} w_j = 1$ and from (2.6) we have then that

$$(q_s, q_t) = \delta_{st} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \qquad s; t = 1, 2; \cdots, k.$$

Consequently the k vectors q_m, $m = 1, 2, \cdots, k$ are orthonormal and span a k-dimensional space. Hence the matrix Q must be an orthogonal matrix. Thus its columns c_s, $s = 1, 2, \cdots, k$ must be orthonormal, i.e., $(c_s, c_t) = \delta_{st}$, $s, t = 1, 2, \cdots, k$.

Consequently we have that

$$\sum_{m=1}^{k-1} w_s \cdot 1 + w_s = 1 = (c_s, c_s)$$

so

$$w_s = \frac{1}{k}, \qquad s = 1, 2, \cdots, k. \tag{2.7}$$

Further

$$\sqrt{w_s w_t} \sum_{m=1}^{k-1} e^{im(\theta_s - \theta_t)} + \sqrt{w_s w_t} = 0 = (c_s, c_t) \qquad s \neq t.$$

Thus $\sum_{m=1}^{k-1} e^{im(\theta_s - \theta_t)} + 1 = 0$. Summing this geometrical progression we obtain that

$$[e^{i(\theta_s - \theta_t)}]^k = 1 \qquad s \neq t = 1, 2, \cdots, k. \tag{2.8}$$

Thus the k angles θ_j are equally spaced as the k^{th} roots of unity. However the k^{th} roots of unity determine the k vertices of a regular polygon inscribed in the unit circle. Thus the k points x_j, $j = 1, 2, \cdots, k$ form a neighbor set about x_0 with the property that these points are the vertices of a regular inscribed polygon of k sides with center x_0 and radius r. With this result and (2.7) the conclusion of the theorem is obtained.

The converse follows at once since (2.6) will be valid for $m = 1, 2, \cdots, k - 1$.

THEOREM 2.2. *In a two dimensional problem a harmonic approximation $\bar{U}_k(x_0, r, v, w)$ which uses equal weights,* (i.e., $w_j = 1/k$, $j = 1, 2, \cdots, k$), *is optimum if and only if the distances r_j and the orientation of the vectors v_j, $j = 1, 2, \cdots, k$, which determine the neighbor set of x_0 are arranged such that $r_j = \hat{r}$ for $j = 1, 2, \cdots, k$ and such that the k points of the neighbor set are the vertices of a regular inscribed polygon of k sides with center x_0 and radius $r = \hat{r}$. The optimum order of approximation is k.*

PROOF. Some of the details will be similar to those of the proof of theorem 2.1. We now have in place of (2.5) the following representation of the operator, i.e.,

$$\bar{U}_k(x_0, r, v, w) = u(x_0) + \frac{1}{k} \sum_{m=1}^{\infty} c_m \left(\sum_{j=1}^{k} r_j^m e^{im\theta_j} \right). \tag{2.9}$$

We will now show then by a proper choice of the θ_j's and r_j's, $j = 1, 2, \cdots, k$, that the terms involving r^m, $m = 1, 2, \cdots, k - 1$ in (2.9) can be made to vanish simultaneously, i.e., we can have

$$\sum_{j=1}^{k} r_j^m e^{im\theta_j} = 0, \qquad \text{for } m = 1, 2, \cdots, k - 1. \tag{2.10}$$

Hence as in the proof of theorem 2.1 if (2.10) holds for $m = 1, 2, \cdots, k - 1$ then the k-point harmonic approximation is optimum and the optimum order of approximation is k. It will turn out that $\hat{r} = r_j$, $j = 1, 2, \cdots, k$.

Consider $t_j = r_j e^{i\theta_j}$, $j = 1, 2, \cdots, k$, then the k functions t_j are such that their first $(k - 1)$ elementary symmetric functions are zero. This follows by appealing to (2.10) for $m = 1, 2, \cdots, k - 1$, and by recursive use of Newton's formula for sums of powers (see, for example, Uspensky [32], chapter XI) for

$$S_l = \gamma_1 S_{l-1} + \gamma_2 S_{l-2} + \cdots + (-1)^{l-1} \gamma_{k-1} S_1 + (-1)^l l \gamma_k = 0$$

where $l < k$ and where S_l is the sum of the l^{th} powers and γ_p is the p^{th} elementary symmetric function. Hence the t_j's are roots of the equation of the form $t^k - c = 0$. But $c = (-1)^{k-1}\prod_{j=1}^{k} t_j$. Hence $r_j = r$ and the $k\theta_j$'s are arranged as the k^{th} roots of unity. Then this part of the proof follows by the same arguments used in theorem 2.1. Similarly the converse follows at once since (2.10) is valid for $m = 1, 2, \cdots, k - 1$.

Having studied two classes of operators relative to neighbor sets we next turn our attention to the operators defined over networks of points which are developed from these neighbor sets.

THEOREM 2.3. *In a two dimensional problem a harmonic approximation obtained from the general operator* $\overline{U}_k(x_0, r, v, w)$ *which uses* $r_j = r$ *or* $w_j = 1/k$ *for* $j = 1, 2, \cdots, k$ *possesses three classes of admissible networks for the two dimensional Dirichlet problem. These are the three regular plane tesselations.*

PROOF. To have an optimum harmonic approximation of order k, $k \geqq 3$, at a given point, say x_0, for the two dimensional case we know from theorems 2.1 and 2.2 that if the operator $\overline{U}_k(x_0, r, v, w)$ has $r_j = r$ or $w_j = 1/k$, $j = 1, 2, \cdots, k$, it is necessary and sufficient that the neighbor set about the point x_0 consists of k points where the k points are the vertices of a regular k-sided polygon with center x_0 and radius r with $w_j = 1/k$, $j = 1, 2, \cdots, k$. If a network of points is to be admissible for the Dirichlet problem we know from definition 2.9 that the structure of its interior points must admit a uniformly optimum approximation on these interior points. Thus from theorems 2.1 and 2.2 this is the same as insisting that the interior points of the network form a regular plane tesselation, (see, for example, Coxeter [9], page 58). Further from Coxeter we find that in the plane there are only three regular tesselations, namely regular hexagons, squares, and equilateral triangles. Hence under these conditions there are only three classes of admissible networks.

The admissible class of networks selected would depend on whether the order of approximation desired is 3, 4, or 6.

2.4. *The discrete Dirichlet problem in higher dimensions.* For the N-dimensional problem, $N \geqq 3$, the questions of optimum harmonic approximations for the class of harmonic functions and admissible networks for the Dirichlet problem have to be approached in a slightly different manner. In the first place when the dimensionality is greater than two, regular polytopes with k vertices exist for only special values of k. Thus we cannot expect a complete generalization of theorems 2.1 and 2.2. Secondly, for $N \geqq 3$, excluding $N = 4$, there is only one regular N-dimensional tesselation, this being formed from the N-dimensional hyper-cube. For $N = 4$ there is one other regular tesselation, (see Coxeter [9], page 156).

We now study the construction of neighbor sets in N-dimensions.

THEOREM 2.4. *In an* N*-dimensional problem a* k*-point harmonic approximation* $\overline{U}_k(x_0, r, v, w)$ *is a harmonic approximation of order* $0(\hat{r}^s)$ *if and only if the* k *points constituting the neighbor set satisfy the following conditions*

$$\sum_{j=1}^{k} w_j r_j v_{jl} = 0 \qquad \text{where } l = 1, 2, \cdots, N, \tag{2.11}$$

$$\sum_{j=1}^{k} w_j r_j^s v_{jl} v_{js} = c\delta_{ls} \qquad \text{where } l, s = 1, 2, \cdots, N, \tag{2.12}$$

and c *is a positive constant.*

(The w_j's, r_j's, and v_{jl}'s are specified by definition 2.4.)

PROOF. By the definition of the operator $\overline{U}_k(x_0, r, v, w)$ and the fact that u is any function of the class of functions harmonic in D we have a valid sum of Taylor series, namely,

$$\overline{U}_k(x_0, r, v, w) = \sum_{j=1}^{k} w_j u(x_j) = u(x_0) + \sum_{j=1}^{k} w_j \left\{ \sum_{l=1}^{N} \left(r_j v_{jl} \frac{\partial}{\partial x_{0l}} \right) u \Big|_{x=x_0} \right\}$$

$$+ \frac{1}{2} \sum_{j=1}^{k} w_j \left\{ \left(\sum_{l=1}^{N} r_j v_{jl} \frac{\partial}{\partial x_{0l}} \right)^2 u \Big|_{x=x_0} \right\} + 0(\hat{r}^3),$$

where $\hat{r}$ is specified by definition 2.5.

Given conditions (2.11) and (2.12) direct substitution then shows that the operator $\overline{U}_k(x_0, r, v, w)$ is a k-point harmonic approximation of order $0(\hat{r}^3)$.

It follows from inspection of the sum of the Taylor series given above that conditions (2.11) and (2.12) are necessary if $\overline{U}_k(x_0, r, v, w)$ is to be a harmonic approximation to order $0(\hat{r}^3)$ for all functions harmonic in D. For if these conditions are not given then it is possible to find harmonic functions in D for which $\overline{U}_k(x_0, r, v, w)$ is not a harmonic approximation to order $0(\hat{r}^3)$.

THEOREM 2.5. *In an N-dimensional problem a harmonic approximation $\overline{U}_{N+1}(x_0, r, v, w)$ which uses $(N + 1)$ points all at the same distance r from x_0 is an optimum harmonic approximation of order $0(r^3)$ if and only if the weights $w_j = 1/(N + 1)$ for $j = 1, 2, \cdots, N + 1$ and the $(N + 1)$ points constituting the neighbor set are restricted to be the vertices of an N-dimensional regular simplex inscribed in the N-sphere of radius r with center x_0.*

PROOF. If we are to have a harmonic approximation of third order, i.e., $\overline{U}_k(x_0, r, v, w) = u(x_0) + 0(\hat{r}^3)$, for the class of all functions harmonic in D it is necessary and sufficient to have conditions (2.11) and (2.12) of theorem 2.4 satisfied. Here $r_j = r = \hat{r}$ for $j = 1, 2, \cdots, k$.

The problem of demonstrating that we have an optimum approximation of order $0(r^3)$ resolves itself to finding the smallest value of k for which we have an approximation of order $0(r^3)$. Since a neighbor set must utilize at least $(N + 1)$ points we will surely have an optimum harmonic approximation if we can show that when $k = N + 1$ we have an approximation of third order.

Consider the following matrix Q

$$Q = \begin{bmatrix} \sqrt{w_1} v_{11} & \sqrt{w_2} v_{21} & \cdots & \sqrt{w_k} v_{k1} \\ \sqrt{w_1} v_{12} & \sqrt{w_2} v_{22} & \cdots & \sqrt{w_k} v_{k2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sqrt{w_1} v_{1N} & \sqrt{w_2} v_{2N} & \cdots & \sqrt{w_k} v_{kN} \\ \sqrt{w_1} & \sqrt{w_2} & \cdots & \sqrt{w_k} \end{bmatrix}.$$

We now assume that the elements of Q satisfy the conditions (2.11) and (2.12) of theorem 2.4 with $r_j = r = \hat{r}$ for $j = 1, 2, \cdots, k$ so as to have an approximation of third order.

Let q_i denote the i^{th} row of Q. Then by (2.11) the inner product of q_i, $i = 1, 2, \cdots, N$, with q_{N+1} yields that $(q_i, q_{N+1}) = 0$, $i = 1, 2, \cdots, N$. Further by (2.12) $(q_i, q_s) = c/r^2\delta_{is}$, $i, s = 1, 2, \cdots, N$. Hence the rows from a basis for a space of $(N + 1)$ dimensions when $k = N + 1$.

As in theorem 2.1 we assume, without loss of generality, that $w_j > 0$, $j = 1, 2, \cdots, N + 1$. We will now show that the $w_j = 1/(N + 1), j = 1, 2, \cdots, N + 1$, and that the $(N + 1)$ points of the neighbor set must be the vertices of an N-dimensional regular simplex inscribed in the N-sphere of radius r with center x_0.

From (2.12) and that $\sum_{i=1}^{N} v_{ji}^2 = 1, j = 1, 2, \cdots, N + 1$ we have that $c = r^2/N$, consequently

$$\sum_{j=1}^{N+1} w_j v_{jl}^2 = \frac{1}{N} \qquad l = 1, 2, \cdots, N. \tag{2.13}$$

Thus the rows of Q can be normalized to yield the matrix M whose rows are orthonormal and span an $(N + 1)$-space. Thus M is an orthogonal matrix and hence its columns c_s, $s = 1, 2, \cdots, N + 1$, must be orthonormal, i.e.

$$M = \begin{bmatrix} \sqrt{Nw_1}v_{11} & \sqrt{Nw_2}v_{21} & \cdots & \sqrt{Nw_{N+1}}v_{N+1,2} \\ \sqrt{Nw_1}v_{12} & \sqrt{Nw_2}v_{22} & \cdots & \sqrt{Nw_{N+1}}v_{N+1,2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sqrt{Nw_1}v_{1N} & \sqrt{Nw_2}v_{2N} & \cdots & \sqrt{Nw_{N+1}}v_{N+1,N} \\ \sqrt{w_1} & \sqrt{w_2} & \cdots & \sqrt{w_{N+1}} \end{bmatrix}.$$

Consequently

$$(c_i, c_i) = Nw_i \sum_{l=1}^{N} v_{il}^2 + w_i = 1 \quad \text{for} \quad i = 1, 2, \cdots, N + 1.$$

Hence we obtain

$$w_i = \frac{1}{N + 1} \qquad i = 1, 2, \cdots, N + 1. \tag{2.14}$$

Further

$$(c_i, c_s) = N\sqrt{w_i w_s} \sum_{l=1}^{N} v_{il}v_{sl} + \sqrt{w_i w_s} = 0, \text{ for } s \neq i = 1, 2, \cdots, N + 1.$$

Hence we have

$$\sum_{l=1}^{N} v_{il}v_{sl} = -\frac{1}{N} \qquad s \neq i = 1, 2, \cdots, N + 1. \tag{2.15}$$

This requirement is the same as insisting that the cosine of the angle between the vectors v_s and v_i equals $-1/N$ for $s \neq i = 1, 2, \cdots, N + 1$. Hence all the

angles between the vectors emanating from x_0 to the points of the neighbor set x_j, $j = 1, 2, \cdots, N+1$ are equal to $\cos^{-1} 1/N$. Thus these $(N+1)$ points determine the vertices of an N-dimensional regular simplex inscribed in the N-sphere of radius r with center x_0. This fact along with (2.14) stating that $w_j = 1/(N+1), j = 1, 2, \cdots, N+1$, yields the desired conclusion of the theorem.

The converse of the theorem is obtained as follows. If the $(N+1)$ points are the vertices of an N-dimensional regular simplex inscribed in the N-sphere of radius r with center x_0 then the converse follows by direct verification. For example, it follows by showing that $M'M = I = MM'$, where I is the $(N+1)$ by $(N+1)$ identity matrix, since this procedure yields that conditions (2.11) and (2.12) are satisfied.

THEOREM 2.6. *In an N-dimensional problem a harmonic approximation* $\overline{U}_{N+1}(x_0, r, v, w)$ *which uses equal weights,* (i.e., $w_j = 1/(N+1), j = 1, 2, \cdots, N+1$) *is an optimum harmonic approximation of order* $0(\hat{r}^3)$ *if an only if the distances* r_j *and the orientation of the vectors* v_j, $j = 1, 2, \cdots, N+1$, *which determine the neighbor set of* x_0 *are arranged such that* $r_j = \hat{r}$ *for* $j = 1, 2, \cdots, N+1$ *and such that the* $(N+1)$ *points constituting the neighbor set are restricted to be the vertices of an N-dimensional regular simplex inscribed in the N-sphere of radius* $r = \hat{r}$ *with center* x_0.

PROOF. Certain details of the proof are similar to those found in the proof of theorem 2.5. For the sake of brevity we will not repeat these details. In place of the matrix Q we now have R where we now make use of the fact that $w_j = 1/k, j = 1, 2, \cdots, k$.

$$R = \begin{bmatrix} r_1 v_{11} & r_2 v_{21} & \cdots & r_k v_{k1} \\ r_1 v_{12} & r_2 v_{22} & \cdots & r_k v_{k2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ r_1 v_{1N} & r_2 v_{2N} & \cdots & r_k v_{kN} \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

We now assume the elements of R satisfy the conditions (2.11) and (2.12) of theorem 2.4 with $w_j = 1/k, j = 1, 2, \cdots, k$ so as to have an approximation of third order.

Let q_t denote the t^{th} row of R. Then by (2.11) the inner product of q_t, $t = 1, 2, \cdots, N$ with q_{N+1} yields that $(q_t, q_{N+1}) = 0$, $t = 1, 2, \cdots, N$. Further by (2.12) $(q_t, q_s) = (k)^2 c\delta_{ts}$, $t, s = 1, 2, \cdots, N$. Hence the rows form a basis for a space of $(N+1)$ dimensions when $k = N+1$.

We will now show that $r_j = r$ for $j = 1, 2, \cdots, N+1$, where

$$r = \hat{r} = \frac{1}{N+1} \sqrt{\sum_{j=1}^{N+1} r_j^2},$$

and that the $(N + 1)$ points of the neighbor set must be the vertices of an N-dimensional regular simplex inscribed in the N-sphere of radius r with center x_0.

From (2.12) and that $\sum_{i=1}^{N} v_{ji}^2 = 1$, $j = 1, 2, \cdots, N + 1$, we now have that

$$c = \frac{\sum_{j=1}^{N+1} r_j^2}{N(N+1)^2},$$

consequently we obtain

$$\sum_{j=1}^{N+1} r_j^2 v_{jl}^2 = \frac{\sum_{j=1}^{N+1} r_j^2}{N}, \qquad l = 1, 2, \cdots, N. \tag{2.16}$$

Thus the rows of R can be normalized to yield the matrix S whose rows are orthonormal and span an $(N + 1)$-space. Thus S is an orthogonal matrix and hence its columns c_s, $s = 1, 2, \cdots, N + 1$, must be orthonormal, i.e.

$$S = \begin{bmatrix} \frac{\sqrt{N} r_1 v_{11}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} & \frac{\sqrt{N} r_2 v_{21}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} & \cdots & \frac{\sqrt{N} r_{N+1} v_{N+1,1}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} \\ \frac{\sqrt{N} r_1 v_{12}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} & \frac{\sqrt{N} r_2 v_{22}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} & \cdots & \frac{\sqrt{N} r_{N+1} v_{N+1,2}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\sqrt{N} r_1 v_{1N}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} & \frac{\sqrt{N} r_2 v_{2N}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} & \cdots & \frac{\sqrt{N} r_{N+1} v_{N+1,N}}{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}} \\ \frac{1}{\sqrt{N+1}} & \frac{1}{\sqrt{N+1}} & \cdots & \frac{1}{\sqrt{N+1}} \end{bmatrix}.$$

Consequently,

$$(c_t, c_t) = \frac{N \sum_{i=1}^{N} r_i^2 v_{it}^2}{\sum_{j=1}^{N+1} r_j^2} + \frac{1}{N+1} = 1, \qquad t, = 1 = 1, 2, \cdots, N + 1.$$

Hence we obtain

$$r_t = \frac{(\sum_{j=1}^{N+1} r_j^2)^{\frac{1}{2}}}{N+1} = r, \qquad t = 1, 2, \cdots, N + 1. \tag{2.17}$$

Further

$$(c_t, c_s) = \frac{N \sum_{i=1}^{N} r_t r_s v_{it} v_{si}}{\sum_{j=1}^{N+1} r_j^2} + \frac{1}{N+1} = 0, \qquad t \neq s = 1, 2, \cdots, N + 1$$

which with (2.17) yields that

$$\sum_{i=1}^{N} v_{it} v_{is} = -\frac{1}{N}, \qquad t \neq s = 1, 2, \cdots, N + 1. \tag{2.18}$$

The completion of this aspect of the proof and the converse follows by using (2.17) and (2.18) and a discussion similar to the one found in the proof of theorem 2.5.

We have seen, subject to the conditions imposed by theorems 2.1 and 2.2, that in the two-dimensional case we are able to improve the order of the harmonic approximation at a point by properly increasing the number of members constituting the neighbor set. However for $N \geqq 3$ we have an interesting development. If we want to have a harmonic approximation based on the operator $\bar{U}_{N+2}(x_0, r, v, w)$ which is of third order and which uses $(N + 2)$ points instead of $(N + 1)$ points and has its weights all equal, i.e., $w_j = 1/(N + 2)$, $j = 1, 2, \cdots, N + 2$, and $r_j = r, j = 1, 2, \cdots, N + 2$, then we will find that this goal is only attainable when N is even.

THEOREM 2.7. *In an N-dimensional problem, N is odd and equal to or greater than three, a harmonic approximation $\bar{U}_{N+2}(x_0, r, v, w)$ of order $0(r^3)$ which uses a neighbor set of $(N + 2)$ points is not possible when the operator $\bar{U}_{N+2}(x_0, r, v, w)$ utilizes a set of equal weights and a set of equal distances.*

PROOF. The proof is by contradiction, i.e., assume it is possible to form a neighbor set using $(N + 2)$ points, $N \geqq 3$ and odd, and have a harmonic approximation of order $0(r^3)$ when $w_j = 1/(N + 2)$ and $r_j = r$ for $j = 1, 2, \cdots, N + 2$. From theorem 2.4 we know that conditions (2.11) and (2.12) must hold. Using the same approach as has been used in the proofs of theorems 2.5 and 2.6 with $k = (N + 2)$ a condition comparable to (2.13) or (2.16) now appears as $\sum_{j=1}^{N+2} v_{jl}^2 = (N + 2)/N$, $l = 1, 2, \cdots, N$, where here the constant c in condition (2.12) of theorem 2.4 equals r^2/N. Thus the conditions to have an approximation of order $0(r^3)$ are the same as insisting that the first $(N + 1)$ rows of the matrix M, given below, be orthogonal, i.e.

$$M = \begin{bmatrix} \dfrac{v_{11}}{\sqrt{\frac{N+2}{N}}} & \dfrac{v_{21}}{\sqrt{\frac{N+2}{N}}} & \cdots & \dfrac{v_{N+1,1}}{\sqrt{\frac{N+2}{N}}} & \dfrac{v_{N+2,1}}{\sqrt{\frac{N+2}{N}}} \\ \dfrac{v_{12}}{\sqrt{\frac{N+2}{N}}} & \dfrac{v_{22}}{\sqrt{\frac{N+2}{N}}} & \cdots & \dfrac{v_{N+1,2}}{\sqrt{\frac{N+2}{N}}} & \dfrac{v_{N+2,2}}{\sqrt{\frac{N+2}{N}}} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \dfrac{v_{1N}}{\sqrt{\frac{N+2}{N}}} & \dfrac{v_{2N}}{\sqrt{\frac{N+2}{N}}} & \cdots & \dfrac{v_{N+1,N}}{\sqrt{\frac{N+2}{N}}} & \dfrac{v_{N+2,2N}}{\sqrt{\frac{N+2}{N}}} \\ \dfrac{1}{\sqrt{N+2}} & \dfrac{1}{\sqrt{N+2}} & \cdots & \dfrac{1}{\sqrt{N+2}} & \dfrac{1}{\sqrt{N+2}} \\ m_{1,N+2} & m_{2,N+2} & \cdots & m_{N+1,N+2} & m_{N+2,N+2} \end{bmatrix}.$$

Further by inspection we see that the first $(N + 1)$ rows q_s, $s = 1, 2, \cdots, N + 1$, of M are of unit length. Consequently we have $(N + 1)$ ortho-

normal vectors q_s in an $(N+2)$-dimensional space. Hence we can find a vector q_{N+2} such that it can be adjoined to the first $(N+1)$ orthonormal vectors so as to have an orthonormal basis for the $(N+2)$-dimensional space. As a result we have that $(q_{N+2}, q_s) = \delta_{N+2,s}$, $s = 1, 2, \cdots, N+2$. Let q_{N+2} be the last row of M, i.e., $q_{N+2} = (m_{1,N+2}, m_{2,N+2}, \cdots, m_{N+2,N+2})$. Hence M is now an orthogonal matrix so its columns are normalized too, i.e., of unit length.

Consider any column, say the s^{th} column c_s. Then

$$(c_s, c_s) = 1 = \frac{1}{(N+2)/N} + \frac{1}{N+2} + m_{s,N+2}^2, \qquad s = 1, 2, \cdots, N+2.$$

So $m_{s,N+2} = (\pm)\, 1/\sqrt{N+2}$ (the proper sign to be determined), for $s = 1, 2, \cdots, N+2$.

But $q_{N+2} = (m_{1,N+2}, m_{2,N+2}, \cdots, m_{N+2,N+2})$ must be orthogonal to the other vectors and in particular to q_{N+1}. However $(N+2)$ is odd since by hypothesis N is odd. It then follows that no matter how the signs are selected for the elements $m_{i,N+2}$, $i = 1, 2, \cdots, N+2$, $(q_{N+2}, q_{N+1}) \neq 0$. But this leads to a contradiction of the assumption that $(N+2)$ points could be found so as to have an approximation of $0(r^3)$. Thus the proof is completed.

When N is even theorem 2.7 can be replaced by a positive statement. The following theorem is useful in describing the structure of the appropriate neighbor set for the operator $\overline{U}_{N+2}(x_0, r, v, w)$.

THEOREM 2.8. *Given $(m+1)$ equal unit vectors such that the angle between any two of these vectors equals $\cos^{-1} 1/m$ then the $(m+1)$ vectors determine a space of m dimension through the origin of those vectors.*

PROOF. Let $v_1, v_2, \cdots, v_{m+1}$ be the given vectors. Consider the Grammian, i.e., the following $(m+1) \times (m+1)$ determinant G.

$$\det G = \begin{vmatrix} (v_1, v_1) & (v_1, v_2) & \cdots & (v_1, v_{m+1}) \\ (v_2, v_1) & (v_2, v_2) & \cdots & (v_2, v_{m+1}) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (v_{m+1}, v_1) & (v_{m+1}, v_2) & \cdots & (v_{m+1}, v_{m+1}) \end{vmatrix}$$

$$= \begin{vmatrix} 1 & -1/m & -1/m & \cdots & -1/m & -1/m \\ -1/m & 1 & -1/m & \cdots & -1/m & -1/m \\ -1/m & -1/m & 1 & \cdots & -1/m & -1/m \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1/m & -1/m & -1/m & \cdots & 1 & -1/m \\ -1/m & -1/m & -1/m & \cdots & -1/m & 1 \end{vmatrix}.$$

Then by subtracting the first row from each of the others and next adding each column to the first we have

$$\det G = \begin{vmatrix} 0 & -1/m & -1/m & \cdots & -1/m & -1/m \\ 0 & 1+1/m & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1+1/m & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ 0 & 0 & 0 & & 1+1/m & 0 \\ 0 & 0 & 0 & & 0 & 1+1/m \end{vmatrix}.$$

Thus the $(m+1)$ vectors v_i, $i = 1, 2, \cdots, m+1$, are linearly dependent, (see Courant and Hilbert [7], page 29). However by inspection the rank of the Grammian is m so the vectors form an m-space. Hence the proof is completed.

This result is in agreement with known geometrical results that no regular simplex of $(m+1)$ vertices exists in less than m dimensions.

THEOREM 2.9. *In an N-dimensional problem, N even, it is possible to have a harmonic approximation* $\overline{U}_{N+2}(x_0, r, v, w)$ *of order* $0(r^3)$ *which uses a neighbor set of* $(N+2)$ *points when the operator* $\overline{U}_{N+2}(x_0, r, v, w)$ *utilizes a set of equal weights and a set of equal distances.*

PROOF. Let v_{n+1} and v_{N+2} be two vectors defined in an $(N+2)$ dimensional space where N is even. The components of the two vectors are given as follows:

$$\text{for } v_{N+1} \text{ they are } g_{N+1,l} = \frac{1}{\sqrt{N+2}} \qquad l = 1, 2, \cdots, N+2$$

$$\text{for } v_{N+2} \text{ they are } g_{N+2,l} = \frac{1}{\sqrt{N+2}} \qquad l = 1, 2, \cdots, \frac{N}{2}+1$$

$$= -\frac{1}{\sqrt{N+2}} \qquad l = \frac{N}{2}+2, \frac{N}{2}+3, \cdots, N+2.$$

Clearly $(v_{N+1}, v_{N+2}) = 0$. It is then possible to find N other vectors v_j, $j = 1, 2, \cdots, N$, with components g_{jl}, $l = 1, 2, \cdots, N+2$, such that together with v_{N+1} and v_{N+2} the $(N+2)$ vectors form an orthonormal basis for $(N+2)$-space. Consequently the following $(N+2) \times (N+2)$ matrix G is orthogonal, i.e.

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1,N/2+1} & g_{1,N/2+2} & g_{1,N/2+3} & \cdots & g_{1,N+2} \\ g_{21} & g_{22} & \cdots & g_{2,N/2+1} & g_{2,N/2+2} & g_{2,N/2+3} & \cdots & g_{2,N+2} \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdots & \cdot \\ g_{N1} & g_{N2} & \cdots & g_{N,N/2+1} & g_{N,N/2+2} & g_{N,N/2+3} & \cdots & g_{N,N+2} \\ \frac{1}{\sqrt{N+2}} & \frac{1}{\sqrt{N+2}} & \cdots & \frac{1}{\sqrt{N+2}} & \frac{1}{\sqrt{N+2}} & \frac{1}{\sqrt{N+2}} & \cdots & \frac{1}{\sqrt{N+2}} \\ \frac{1}{\sqrt{N+2}} & \frac{1}{\sqrt{N+2}} & \cdots & \frac{1}{\sqrt{N+2}} & \frac{-1}{\sqrt{N+2}} & \frac{-1}{\sqrt{N+2}} & \cdots & \frac{-1}{\sqrt{N+2}} \end{bmatrix}.$$

Since G is orthogonal, the rows have the following properties

$$\sum_{l=1}^{N+2} g_{jl} = 0, \qquad j = 1, 2, \cdots, N, \tag{2.19}$$

since $(v_j, v_{N+1}) = 0$

$$\sum_{l=1}^{N+2} g_{jl}g_{kl} = \delta_{jk}, \qquad j, k, = 1, 2, \cdots, N. \tag{2.20}$$

Likewise the columns of G must satisfy the following properties.

$$(c_l, c_s) = \delta_{ls}, \qquad l, s = 1, 2, \cdots, N+2 \tag{2.21}$$

$$(c_l, c_l) = 1 = \sum_{j=1}^{N} g_{jl}^2 + \frac{2}{N+2}, \quad l = 1, 2, \cdots, N+2 \tag{2.22}$$

Further

$$(c_l, c_s) = 0 = \sum_{j=1}^{N} g_{jl}g_{js} + \frac{2}{N+2} \text{ for}$$

$$\begin{cases} l \neq s = 1, 2, \cdots, \dfrac{N}{2} + 1 \\ l \neq s = \dfrac{N}{2} + 2, \dfrac{N}{2} + 3, \cdots, N+2 \end{cases} \tag{2.23}$$

and

$$(c_l, c_s) = 0 = \sum_{j=1}^{N} g_{jl}g_{js} + 0 \text{ for} \begin{cases} l = 1, 2, \cdots, \dfrac{N}{2} + 1 \\ s = \dfrac{N}{2} + 2, \dfrac{N}{2} + 3, \cdots, N+2. \end{cases} \tag{2.24}$$

Now consider the following $(N + 2)$ vectors f_i, $i = 1, 2, \cdots, N + 2$ defined in N space whose components are formed from the first N elements of each column of G. When representing these vectors we have the following unavoidable perversity of the row-column notation, i.e., note the order of the indices in the expression of f_{is} given below. Thus $f_i = (f_{i1}, f_{i2}, \cdots, f_{iN})$, $i = 1, 2, \cdots, N + 2$, where

$$f_{is} = \sqrt{\frac{N+2}{N}}\, g_{si}, \qquad s = 1, 2, \cdots, N.$$

From (2.22) the vectors f_i, $i = 1, 2, \cdots, N + 2$, are unit vectors. From (2.19) we have

$$\sum_{i=1}^{N+2} f_{is} = 0, \qquad s = 1, 2, \cdots, N. \tag{2.25}$$

Likewise from (2.20)

$$\sum_{i=1}^{N+2} f_{is}f_{it} = c\delta_{st}, \qquad s, t = 1, 2, \cdots, N, c = \frac{N+2}{N}. \tag{2.26}$$

Thus the unit vectors f_i, $i = 1, 2, \cdots, N + 2$, determine $(N + 2)$ points on the unit N-sphere about the origin. The conditions (2.25) and (2.26) specify

precisely the conditions needed for theorem 2.4 to be valid. Thus these $(N + 2)$ points form a neighbor set so that $\overline{U}_{N+2}(0, 1, v, w)$ is a harmonic approximation of third order. Hence the more general case for $\overline{U}_{N+2}(x_0, r, v, w)$ follows at once. Thus the proof is completed.

Description of the appropriate neighbor set: We will now describe how the $(N + 2)$ points of theorem 2.9 are arranged on the unit sphere with center at the origin. The description can be immediately generalized for the case of an arbitrary radius r and center x_0.

Consider the first $(N + 2)/2$ vectors f_i, $i = 1, 2, \cdots, \frac{1}{2}N + 1$. From (2.23) the angle between any two of these unit vectors equals $\cos^{-1}(2/N)$. Likewise, the angle between any two unit vectors of the second set, f_i, $i = N/2 + 2$, $N/2 + 3, \cdots, N + 2$, equals $\cos^{-1}(2/N)$. From theorem 2.8 the first $(N/2 + 1)$ vectors f_i determine an $(N/2)$-space through the origin. Likewise the last $(N/2 + 1)$ vectors determine an $(N/2)$-space through the origin. By condition (2.24) every vector of the first set, i.e., f_i, $i = 1, 2, \cdots, N/2 + 1$, is orthogonal to every vector of the second set, and conversely. Thus these two spaces of dimension $(N/2)$ are orthogonal to each other. Hence the $(N + 2)$ points of the neighbor set of the N space can be looked upon as being formed from two separate sets of $(N/2 + 1)$ points belonging to orthogonal subspaces of $(N/2)$ dimensions. Hence the description is complete. In each $(N/2)$-subspace the corresponding $(N/2 + 1)$ points determine a regular simplex since the corresponding unit vectors f_i that determine these points have the property, due to (2.23), that the angle between any two of the vectors of the $(N/2)$ subspace is equal to $\cos^{-1}(2/N)$.

Without the restriction that N be even theorem 2.9, its method of proof, and the description of the appropriate neighbor set can be extended so that a harmonic approximation of order $0(r^3)$ exists whenever the neighbor set consists of k points, k is even, and in addition $N + 3 \leqq k \leqq 2N$. However a neighbor set using this many points will not be optimum for a harmonic approximation of order $0(r^3)$ since this order of approximation can be obtained using $k = N + 1$. For this reason and because the proof would have been increased in size without contributing more information relevant to the discussion theorem 2.9 was phrased only for $k = N + 2$, N even. Thus we next consider what general conditions should be specified to have a harmonic approximation of order $0(\dot{r}^4)$.

THEOREM 2.10. *In N-dimensions a k-point harmonic approximation $\bar{U}_k(x_0, r, v, w)$ is a harmonic approximation of order $0(\dot{r}^4)$ if and only if the k points constituting the neighbor set, in addition to the conditions* (2.11) *and* (2.12) *of theorem* 2.4, *satisfy the following,*

$$\sum_{j=1}^{k} w_j r_j^3 (v_{jl}^3 - 3v_{js}v_{jl}^2) = 0, \quad l \neq s = 1, 2, \cdots, N, \tag{2.27}$$

and

$$\sum_{j=1}^{k} w_j r_j v_{js} v_{jt} = 0, \quad l \neq s \neq t = 1, 2, \cdots, N. \tag{2.28}$$

PROOF. This proof essentially follows the arguments used in the proof of theorem 2.4. However it is necessary to study one additional term in the sum

of the Taylor series expansion for $\bar{U}_k(x_0, r, v, w)$, namely

$$\frac{1}{3!}\sum_{j=1}^{k} w_j \left\{ \left(\sum_{l=1}^{N} r_j v_{jl} \frac{\partial}{\partial x_{0l}} \right)^3 u \Big|_{x=x_0} \right\}.$$

Further, since the class of functions being considered is harmonic in D the following additional fact is used;

$$\frac{\partial^3 u(x)}{\partial x_{0l}^3} = -\sum_{i \neq l=1}^{N} \frac{\partial^3 u(x)}{\partial x_{0i} \partial x_{0l}^2}, \qquad l = 1, 2, \cdots, N.$$

THEOREM 2.11. *In N-dimensions a k-point harmonic approximation $\bar{U}_k(x_0, r, v, w)$ is a harmonic approximation of order $0(r^4)$ if $k = 2N$, the set of weights $w_j = 1/2N$, $j = 1, \cdots, 2N$, and the neighbor set of x_0 is arranged such that the $2N$ points which constitute the neighbor set of x_0 are also neighbors of x_0 on an N-dimensional lattice which includes x_0 and whose basic cell is an N-dimensional hyper-cube with neighbor length equal to r.*

PROOF. If the neighbor set of x_0 consists of $2N$ points and these points are the $2N$ neighbors of x_0 on the N-dimensional lattice whose basic cell is the N-dimensional hyper-cube then the conditions of theorem 2.10 can be verified. Since this result is well known for $N = 3$ and since the actual details of the verification are easy to carry out they will not be considered here.

As mentioned in section 2.1 the network of points specified in theorem 2.11 has been considered for some time. However no attention has been given to determine whether this is an admissible network for the Dirichlet problem, i.e., the network when viewed from its neighbor sets yields a uniformly optimum harmonic approximation of order $0(r^4)$ on the interior points. From theorem 2.3 we know this is an admissible network when $N = 2$.

It is clear that for the three dimensional case an approximation of order $0(r^4)$ cannot be achieved with a neighbor set of less than $2N = 6$ points when the weights $w_j = 1/2N$ and distances $r_j = r$ for $j = 1, 2, \cdots, 2N$. This last remark can be seen from the following reasoning. Though a neighbor set of four points can serve for a harmonic approximation of order $0(r^3)$ it is not possible to achieve a harmonic approximation of order $0(r^4)$ with four points for the class of all functions harmonic in D. From theorem 2.7 we know that a neighbor set of five points will not even yield a harmonic approximation of order $0(r^3)$. Thus the need for $2N = 6$ points. There still remains the question of arrangement of the six points.

For a desired harmonic approximation of order $0(r^4)$ with $N = 4$ the possible existence of, as well as the appropriate arrangement of, a neighbor set consisting of less than $2N$ points is still an open question. Further, the more general formulation of specifying only that the weights or the distances are equal still requires more study. It has not been possible to determine what the proper criteria are so as to have an optimum harmonic approximation of order $0(r^4)$.

For those situations where we specify the weights $w_j = 1/2N$ and the distance $r_j = r$ for $j = 1, 2, \cdots, 2N$ the following *conjecture* seems most reasonable: The converse of theorem 2.11 is valid, i.e., in N-dimensions the harmonic ap-

proximation $\bar{U}_{2N}(x_0, r, v, w)$ is an optimum harmonic approximation of order $0(r^4)$ only if its neighbor set consists of the $2N$ points which are the neighbors of x_0 on an N-dimensional lattice which includes x_0 and whose basic cell is an N-dimensional hyper-cube with neighbor length equal to r. Clearly if this conjecture is valid then this network of points forms an admissible network for the Dirichlet problem.

We will now show it is possible to find an admissible network for the Dirichlet problem when the order of the approximation is specified to be $0(r^3)$.

The "N-dimensional neighbor simplex network" about to be described is obtained from the alternate use of points of two parallel lattices. Each interior point of the network will have a neighbor set of $(N + 1)$ points. These $(N + 1)$ points belong to the network and they form a regular N-dimensional simplex of radius r with center the point in question. Each interior point belongs to one of the two lattices while its neighbor set belongs to the other lattice. The most expedient description utilizes the embedding of the N-dimensional network in $(N + 1)$-space.

DEFINITION 2.10. *N-dimensional neighbor simplex network.* This network is defined in an $(N + 1)$-dimensional space. The integral coordinates $(y_1, y_2, \cdots, y_{N+1})$ of each point y of the network satisfy the following conditions[3])

$$y_1 + y_2 + \cdots + y_{N+1} = 0 \tag{2.29}$$

$$y_1 \equiv y_2 \equiv \cdots \equiv y_N \bmod N + 1 \tag{2.30}$$

and either

$$y_1 \equiv 0 \bmod N + 1 \tag{2.31a}$$

or

$$y_1 \equiv 1 \bmod N + 1 \tag{2.31b}$$

THEOREM 2.12. *The N-dimensional neighbor simplex network is an admissible network for the N-dimensional Dirichlet problem when using a harmonic approximation obtained from the general operator $\bar{U}_{N+1}(x_0, r, v, w)$ which uses $r_j = r$ or $w_j = 1/(N + 1)$, for $j = 1, 2, \cdots, N + 1$.*

PROOF. From theorems 2.5 and 2.6 it is possible to have an optimum harmonic approximation of order $0(r^3)$ at a point, say x_0, belonging to the N-dimensional domain D if and only if the neighbor set of x_0 consists of the $(N + 1)$ points which are the vertices of an N-dimensional regular simplex with center x_0 and radius r with $w_j = 1/(N + 1)$ for $j = 1, 2, \cdots, N + 1$. Definition 2.10 specifies precisely that the N-dimensional neighbor simplex network is a network of such neighbor sets for the interior points. Thus using the neighbor simplex network it is possible to have a uniformly optimum harmonic approximation of order $0(r^3)$ in the sense of definition 2.7. Hence the neighbor simplex network is an admissible network for the Dirichlet problem.

[3] This approach was kindly suggested by R. Steinberg. Motzkin [25] has since mentioned other possible descriptions.

As pointed out by T. S. Motzkin the description of the neighbor simplex network given in definition 2.10 is readily adapted to serve as a program for a general purpose computer in the following way.

DEFINITION 2.11. *Program for the neighbor simplex network:*

1. Let $y = (y_1, \cdots, y_{N+1})$ be any point of the network of definition 2.10. Depending on whether it satisfies condition (2.31a) or (2.31b) call is an a- or b- point.

2. Let $e = \sum_{j=1}^{N+1} e_j$ where $e_j = (0, 0, 1, 0, \cdots, 0)$, i.e., the unit appears in the j^{th} coordinate.

3. The $N + 1$ neighbors for y are given as follows: either

a. $a' = y + e - (N + 1)e_j$, $j = 1, 2, \cdots, N + 1$, if y is an a-point of the network, or

b. $b' = y - e + (N + 1)e_j$, $j = 1, 2, \cdots, N + 1$, if y is a b-point of the network.

King [19], [20] has successfully employed a network (which he calls a "tetrahahedral lattice") in the study of High Polymers. He credits this network to Mark and Tobolsky. This approach is very well suited to serve as another method of preparing a program for a neighbor simplex network in the 3-dimensional Dirichlet problem. For the 3-dimensional problem it represents a more simple description than the one given here. It is obtained by a projection onto a different 3-dimensional hyperplane. However the approach will not generalize to yield a program for all N.

2.5. *Behavior near the boundary.* One of the most difficult challenges confronting the numerical analyst who uses discrete approximations based on predetermined networks of points for solving boundary value problems is the question of how the boundary $\Gamma[D]$ of the given domain is to be replaced by a discrete set of points. Under most circumstances, i.e., unless the boundary $\Gamma[D]$ is of a very special nature such as the sum of surfaces of hyper-cubes, the order to which the difference operator approximates the differential operator will not be maintained for points near the boundary $\Gamma[D]$, i.e., at the intermediate points, unless an interpolation method is introduced. Wasow [36] contains a review of the literature on this topic.

Intuitively one would feel that the order of the *discretization error*, i.e., the order of the error between the unknown solution and the solution of the discrete approximation, would depend on the particular interpolation method being employed. As intuitively obvious as this observation is, fully satisfactory mathematical studies of this question are still lacking.

Briefly the weakness of existing studies may be summarized into three categories: 1) the studies have required additional assumptions concerning regularly either of the boundary or of the unknown solution or both, 2) the studies have been a study in the large, 3) the studies using just the numerical data alone are too incomplete.

Clearly what is desired is a statement to the effect that if the differential operator is being approximated to a given order then the discretization error is of this same order.

Assuming that the third order partial derivatives of the unknown solution are bounded up to within a small strip of the boundary $\Gamma[D]$ it is then possible to demonstrate that the discrete Dirichlet problem can be solved on the neighbor simplex network of definition 2.10 and that the order of the discretization error is the same throughout the entire domain without using an interpolation scheme near the boundary. At the boundary one simply replaces the boundary by the discrete set of points where the line segments joining neighbors of the neighbor simplex intersect the original boundary $\Gamma[D]$. The details of the verification of this material will not be given since the details, though long, follow by direct application of the material found in Wasow [36], i.e., it is only necessary to appeal to the material by Collatz [4] and Gerschgorin [13] in the manner that has been done by Wasow for the case when the interior points form a lattice whose basic cell is an N-dimensional hyper-cube. However, here one will only need to assume that the third order partial derivatives of the unknown solution are bounded rather than the fourth order partial derivatives.

2.6. *A comparison between two classes of networks.* It has been customary to suggest that the network of points to be used, at least for the interior points of the network, should form an N-dimensional lattice whose basic cell is an N-dimensional hyper-cube. Unless some sort of an elaborate interpolation scheme is used near the boundary of the given domain D the finite difference technique using this lattice will only approximate the unknown solution of the Dirichlet problem to order $0(\hat{r})$, e.g., see Collatz [4], Gerschgorin [13], Wasow [36]. With this lattice each interior point of the lattice has $2N$ neighbors.

Using the N-dimensional neighbor simplex discussed in sections 2.4 and 2.5 it is possible to have an approximation of order $0(\hat{r})$ where each interior point needs only $(N + 1)$ neighbors and an elaborate interpolation scheme will not be needed near the boundary, (see section 2.5).

Since both classes of networks will have to be programmed for a computer, it may develop that a very significant saving in machine time can be achieved by using an N-dimensional simplex network if an approximation of order three is satisfactory. We next turn our attention to see that a Monte Carlo estimate for the Dirichlet problem can be obtained by utilizing a network of this type.

3. Random walks using closed neighbor simplex networks. 3.1. *Introduction.* Most of the literature on Monte Carlo methods for boundary value problems has been concerned with statistical estimates of the unknown solution of an approximate finite difference problem that approximates the given boundary value problem. This literature includes the relevant work by Polya [28], Phillips and Wiener [27], Courant, Friedrichs, and Lewy [6], Lüneburg [22], Petrowsky [26], Khintchine [21], Curtiss [10], and Wasow [35].

Hence there are two inherent sources of error in a discrete Monte Carlo method. One error originates from the replacement of the original boundary value problem by a discrete approximation. The second error results from the statistical fluctuation of the statistic being used to estimate the solution of the discrete approximation. Some attention to the first source of error has been

given in the previous sections. Since the second source of error is almost an independent problem in Statistical Design Theory (e.g. along the lines considered by Kahn and Marshall [16]) we will limit attention here to showing that a Monte Carlo estimate for the N-dimensional Dirichlet problem can be obtained by using random walks on the closed neighbor simplex networks. It will be seen that when considering random walks on the closed neighbor simplex networks it will only be necessary to consider stochastic models which have discrete parameters. Further it will be seen that the model will specify the position of any future state of the process to be one of a finite number of predetermined points. Therefore subject to these conditions only Markov chains (see Doob [11] or Feller [12] for a definition) need be utilized. Consequently it is possible to make use of the earlier work by Petrowsky [26] and his predecessors. We next consider replacing the given domain.

A replacement of a domain D and its boundary $\Gamma[D]$ of the N-dimensional Dirichlet problem by an N-dimensional closed simplex network is obtained in the following manner.

About a given point where a solution to the Dirichlet problem is desired one replaces the given domain D by constructing an N-dimensional neighbor simplex network (see definition 2.10) such that the given point belongs to the network. The boundary $\Gamma[D]$ of the domain D is replaced by the discrete set of points formed from where the line segments joining neighbors of the given neighbor simplex network intersect the boundary $\Gamma[D]$. These points will be called the *discrete boundary* or the *boundary points of the network*. The collection of points consisting of interior, intermediate, and boundary points of an N-dimensional neighbor simplex relative to a given domain D and its boundary $\Gamma[D]$ will be called an *N-dimensional closed neighbor simplex network* relative to $D + \Gamma[D]$.

3.2. *The random walk procedure.* It is possible to trace the introduction and use of the term random walks or random flights to the study on vibrations made by Lord Rayleigh and reported in the Philosophical Magazine of 1880. The use of the term "random walk" has been so expanded by some authors that it now includes the description of the movement of any object or particle that makes a countable sequence of transitions on a point set in a Euclidean space in a manner that can be described by a stochastic process.

In this paper a random walk will mean that a particle may move on a predetermined network of points. Further, the movement of the particle is specified by a particular type of Markov chain. A complete detailed description of a random walk on a closed neighbor simplex network will not be given since only minor modifications are needed in order to describe a walk of this type in terms of the existing literature on random walks such as found in Curtiss [10] or Feller [12]. Briefly, a random walk on an N-dimensional closed neighbor simplex network means that at any given state of the process one of the following three situations applies.

1. The particle is at an interior point of the network and a single transition will find the particle moving with equal probability to a point of the closed neighbor simplex network which is one of the $(N + 1)$ members of the neighbor set of this given interior point.

2. The particle is at an intermediate point (see definition 2.8) of the network and a single transition will find the particle moving with equal probability to a point of the closed neighbor simplex network which is either one of the interior members of the neighbor set of this given intermediate point or one of the boundary points formed from this intermediate point.

3. The particle is at a boundary point and consequently this walk will be terminated.

Note: As is customary a step of the random walk on the closed neighbor simplex network is simply a single transition of the process and the length of the step is just r, the radius of the neighbor set.

3.3. *Statistical estimation of the discrete approximation.* If a random walk that has originated from a point of interest terminates, say at the point of y, i.e., the random walk reaches the point y of the boundary and the walk is stopped, then the value of the boundary function $f(x)$ at the boundary y is called the *tally of the random walk*. The *expected value* $E[x_0 ; \hat{r})$ will denote the expected value of the tallies of the random walks taking place on a predetermined network and originating at x_0, i.e., $E[x_0 ; \hat{r}] = \sum_y f(y)Pr[y \mid x_0 ; \hat{r}]$, where $Pr[y \mid x_0 ; \hat{r}]$ is the probability that the random walk relative to a given closed network with neighbor sets of radius $\hat{r}$ terminates at y having originated at x_0, the summation on y is extended over the discrete set of boundary points.

Note: From theorem 3.1 it will be seen that $Pr[y \mid x_0 ; r]$ is the discrete analog of the normal derivative of Green's function, (see also Tamarkin and Feller [30], page 183).

We will now refer to the work by Petrowsky [26] to demonstrate, subject to certain restrictions, that the random walks on the closed neighbor simplex network yield valid approximations for the solution to the Dirichlet problem. Though Petrowsky's work has certain limitations it is being used here since it is still the most complete on discrete approximations for the Dirichlet problem. In addition to requiring that $f(x)$ be continuous on $\Gamma[D]$ he also requires (this is not a severe requirement) that $f(x)$ be uniformly bounded in the complement of D and that $f(x)$ have no surfaces of discontinuities in common with $f(x)$ on $\Gamma[D]$, (the N-dimensional generalization of Petrowsky's work on pages 427 and 438). However in three or more dimensions his method of proof makes it necessary to require that the boundary surface $\Gamma[D]$ consist of a finite number of surfaces where the functions specifying these surfaces are of class C^2, i.e. these functions possess continuous second order partial derivatives.

THEOREM 3.1. *Subject to the above mentioned regularity conditions imposed on* $f(x)$ *and* $\Gamma[D]$ *the expected value* $E[x_0 ; r]$ *of the tallies of the random walks on the closed neighbor simplex network converges to the solution of the Dirichlet problem at the point* x_0 *as the length of the step* r *tends to zero.*

Proof. The extensive details of the proof will not be given because subject to the conditions imposed on $f(x)$ and $\Gamma[D]$ the proof follows by direct application of the material by Petrowsky [26]. One essentially has to show that the first two moments and the third absolute moment of the transition distribution of the random walks on the closed neighbor simplex networks have the proper

asymptotic behavior as the step length of the random walk tends to zero. Further it is necessary to show that the walks terminate with probability one.

We will conclude with a few brief comments concerning machine techniques.

3.4. *Machine techniques.* One way to form a general N-dimensional program for a computer to generate a neighbor simplex network has been suggested in definition 2.11. The question of how to select a future position of the walk is a matter of selecting uniformly at random a number $j, j = 1, 2, \cdots, N + 1$, i.e., each value of j has probability $(1/(N + 1))$ of being selected, and carrying out step three of definition 2.11. Teichroew [31] has a detailed discussion of how a machine can be used for computations of this type. The question of how to write a program for a machine so as to store most efficiently the information concerning the discrete set of boundary points would be out of place in this paper. It should be noted however that this last question deserves careful thought. It will occur in any numerical procedure that utilizes a digital machine; it is surely not peculiar to Monte Carlo.

Briefly, a simple machine procedure would involve the following steps. 1) a random walk is started from the point in question, i.e., where the solution is desired, 2) if the walk terminates the boundary value is tallied, 3) the process is repeated enough times to insure the precision desired, 4) the average of the tallies is used as an estimate of the unknown solution (a more refined technique might not use this average).

Note: Though the random walk process will terminate with probability 1, if a machine process is to be economically feasible it may be efficient to have some provision (e.g. truncation procedures) to handle the occurrence of the exceptional walks which are of long duration.

Acknowledgement. The author wishes to express his appreciation to Professors George W. Brown and Robert Steinberg for their generous contributions and encouragement.

REFERENCES

[1] J. Albrecht, "Taylor-Entwicklungen und finite Ausdrücke für Δ und $\Delta\Delta u$," *Z. Angew. Math. Mech.*, v. 33 (1953), pp. 41–48.

[2] E. F. Beckenbach and M. Reade, "Mean-Values and Harmonic Polynomials," *Trans. Amer. Math. Soc.*, v. 53 (1943), pp. 230–238.

[3] W. G. Bickley, "Finite Difference Formulae for the Square Lattice," *Quart. J. Mech. Appl. Math.*, v. 1 (1948), pp. 34–42.

[4] L. Collatz, "Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 13 (1933), pp. 56–57.

[5] ———, *Numerische Behandlung von Differentialgleichungen*, Springer-Verlag, Berlin, 1951.

[6] R. Courant, K. Friedrichs, and H. Lewy, "Über die partiellen Differenzengleichungen der Mathematischen Physik," *Math. Ann.*, v. 100 (1928), pp. 32–74.

[7] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, Vol. I, Springer, 1931.

[8] ———, *Methoden der Mathematischen Physik*, Vol. II, Springer, 1937.

[9] H. S. M. Coxeter, *Regular Polytopes*, Pitman, 1947.

[10] J. H. Curtiss, "Sampling Methods Applied to Differential and Difference Equations," *Seminar on Scientific Computation, November, 1949*, International Business Machines Corp., New York, 1950.
[11] J. L. Doob, *Stochastic Processes*, Wiley, 1952.
[12] W. Feller, *Probability Theory and Its Applications*, Vol. I, Wiley, 1950.
[13] S. Gerschgorin, "Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 10 (1930), pp. 373–382.
[14] L. E. Grinter, *Numerical Methods of Analysis in Engineering*, MacMillan, 1949.
[15] H. A. Heilbroon, "On Discrete Harmonic Functions," *Proc. Cambridge Philos. Soc.*, v. 45 (1949), pp. 194–206.
[16] H. Kahn and A. W. Marshall, "Methods of Reducing Sample Size in Monte Carlo Computations," *J. Operations Res. Soc. Amer.*, v. 1 (1953), pp. 263–278.
[17] O. D. Kellogg, "Recent Progress with the Dirichlet Problem," *Bull. Amer. Math. Soc.*, v. 32 (1926), pp. 601–625.
[18] ———, *Foundations of Potential Theory*, Ungar, 1929.
[19] G. W. King, "Stochastic Methods in Statistical Mechanics," *Monte Carlo Method*, pp. 12–18, U. S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series 12.
[20] ———, *Research in the Physical and Structural Analysis of High Polymers by Punched Card Methods*. Final Report, Little, Inc. Office of Naval Research, Contract N6-onr-228 T. O. II.
[21] A. Khintchine, *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung*, Chelsea, 1948.
[22] R. Lüneburg, "Das Problem der Irrfahrt ohne Richtungsbeschränkung und die Randwertaufgabe der Potentialtheorie," *Math. Ann.*, v. 104 (1934), pp. 700–738.
[23] N. Metropolis and S. Ulam, "The Monte Carlo Method," *J. Amer. Statist. Assoc.*, v. 44 (1949), pp. 335–341.
[24] W. E. Milne, *Numerical Solution of Differential Equations*, Wiley, 1953.
[25] T. S. Motzkin, "Neighbor Sets for Random Walks and Difference Equations," *in manuscript*.
[26] I. Petrowsky, "Uber das Irrfahrtproblem," *Math. Ann.*, v. 109 (1934), pp. 425–444.
[27] H. B. Phillips and N. Wiener, "Nets and the Dirichlet Problem," *J. Math. Physics*, v. 2 (1923), pp. 105–124.
[28] G. Polya, "Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Strassennetz," *Math. Ann.*, v. 84 (1921), pp. 149–160.
[29] R. V. Southwell, *Relaxation Methods in Theoretical Physics*, Oxford, 1946.
[30] J. D. Tamarkin and W. Feller, *Lectures on Partial Differential Equations*, Brown University, 1941.
[31] D. Teichroew, *Distribution Sampling with High Speed Computers*, Ph.D. Thesis, University of North Carolina, 1953.
[32] J. V. Uspensky, *Theory of Equations*, McGraw-Hill, 1948.
[33] C. de La Vallée Poussin, *Les Nouvelles Méthodes de la Théorie du Potentiel et le Probléme Généralisé de Dirichlet*, Actualités Scientifiques et Industrielles 578, Hermann and Cie, 1937.
[34] J. L. Walsh, "A Mean value theorem for polynomials and harmonic polynomials," *Bull. Amer. Math. Soc.*, v. 42 (1936), pp. 923–930.
[35] W. Wasow, "Random Walks and the Eigenvalues of Elliptic Difference Equations," *J. Research Nat. Bur. Standards*, v. 46 (1951), pp. 65–73.
[36] ———, "Discrete Approximations to Elliptic Differential Equations," *in manuscript*.

University of California, Los Angeles
and Cornell University

(Received July 7, 1955)

INVERSE LAPLACE TRANSFORMS EXPRESSED AS NEUMANN SERIES

By Enzo Cambi

1. The problem of inverting a Laplace transform, i.e., of finding the original function $f(t)$, given the transform

$$F(p) = \int_0^\infty e^{-pt} f(t)\, dt$$

is known to be much more complicated than that of determining the direct transform, that is, of determining $F(p)$, given $f(t)$. In the direct case, actually, the definition itself allows a straightforward determination of $F(p)$.

The inversion is substantially equivalent to the solution of an integral equation with the nucleus e^{-pt}. The solution of the equation can be expressed in the form of the integral of $F(p)e^{pt}$ taken along a "vertical" path in the complex p-plane, but it is known that the actual computation of the complex integral is often impractical and may also lead, if the formula is applied carelessly, to fallacious results. For this reason, many other methods of inversion have been developed, none of which can be regarded as completely general.

The method exposed below is also by no means general, but can afford a satisfactory solution in a number of cases, since it expresses $f(t)$ in the form of a series of good convergence: consequently, the interest of the method is chiefly numerical, though, in some cases, the sum of the series obtained can be written in closed form.

2. The procedure is based on the knowledge of the Laplace trasform of a Bessel function of the first kind

$$\mathfrak{L}\{J_n(t)\} = \int_0^\infty e^{-pt} J_n(t)\, dt = \frac{\{\sqrt{1+p^2} - p\}^n}{\sqrt{1+p^2}} \quad [n \geqq 0] \tag{1}$$

If we put $p = \sinh u$ we obtain

$$\int_0^\infty e^{-t\cdot\sinh u} J_n(t)\, dt = \frac{e^{-nu}}{\cosh u};$$

so that, if we define for brevity, as Λ-trasform of a function $f(t)$, the integral

$$\Lambda\{f(t)\} = \cosh u \int_0^\infty e^{-t\cdot\sinh u} f(t)\, dt \tag{2}$$

we can write the last result in the form

$$\Lambda\{J_n(t)\} = e^{-nu}$$

Suppose now that a given $f(t)$ can be expanded around $t = 0$ in a Neumann series, thus:

$$f(t) = a_0 J_0(t) + a_1 J_1(t) + a_2 J_2(t) + \cdots \tag{3}$$

where the a_n are constants. By multiplying the series (3) by $e^{-t \cdot \sinh u}\, dt$, integrating between zero and infinity (assuming that the series may converge uniformly), and finally multiplying by cosh u, we obtain

$$\Lambda\{f(t)\} = a_0 + a_1 e^{-u} + a_2 e^{-2u} + \cdots \tag{4}$$

that is, a series of powers of e^{-u}. Conversely, whenever the Λ- trasform of an unknown function can be expressed as a convergent power series of e^{-u}, the function itself can be expressed as a Neumann series, having the same coefficients as the power series considered.

3. The region of convergence of a Neumann series is the same as that of a power series that may represent the same function; from the point of view of functional analysis, therefore, Neumann series and power series are equivalent. From the standpoint of numerical calculation, however, Neumann series are usually more convenient than power series (provided a table of Bessel Functions of integral order is available) since, for a given t, the Bessel functions $J_n(t)$ constitute a sequence that tends rapidly to zero with increasing n, while the powers t^n represent an increasing sequence, as soon as t is larger than unity. For a given interval of t, the Bessel functions, beyond a certain order n depending on the value of t, decrease so rapidly, that an estimate of the approximation attainable by stopping the expansion at a certain term can often be obtained in advance, if it is known that the coefficients a_n do not show an exceptionally steep rate of increase with increasing n. As a counterpart to the practicality of the procedure, its applicability is of course subject to some obvious limitations. In fact, the assumption that the Λ-trasform of $f(t)$ may be expanded in a convergent series of form (4) implies that the generality of $f(t)$ be considerably restricted.

The existence of a non-zero radius of convergence for (4) implies, for instance, that $\sqrt[n]{a_n}$ tend to a finite limit for n tending to infinity. In this assumption, as the nth term of (3), that is, $J_n(t)$, behaves asymptotically as $t^n/(2^n n!)$, the radius of convergence of (3) becomes infinite in any case, and this excludes the existence of any (finite) singularity in $f(t)$.

4. The Λ-trasform of a function is known immediately as soon as the Laplace trasform is known: it is sufficient to replace p by sinh u and to multiply the result by cosh u. The result will be expressible as a power series of e^{-u} when $f(t)$ is regular in the whole plane.

We shall apply the procedure to the deduction of some known results concerning Neumann series.

The Laplace trasform of sin t is $1/(1 + p^2)$; hence, immediately

$$\Lambda\{\sin t\} = \frac{1}{\cosh u} = \frac{2}{e^u + e^{-u}} = 2(e^{-u} - e^{-3u} + e^{-5u} - \cdots)$$

and $\sin t = 2\{J_1(t) - J_3(t) + J_5(t) - \cdots\}$. Similarly, from $\mathfrak{L}\{\cos t\} = p/(1 + p^2)$, we obtain

$$\Lambda\{\cos t\} = \sinh u/\cosh u = 1 - 2e^{-2u} + 2e^{-4u} - \cdots,$$

so that

$$\cos t = J_0(t) - 2J_2(t) + 2J_4(t) - \cdots .$$

Finally, given $\mathfrak{L}\{t^n\} = n!/p^{n+1}$, we have

$$t^n = n! \frac{\cosh u}{(\sinh u)^{n+1}} = 2^n n! e^{-nu} \left\{ \frac{2}{(1 - e^{-2u})^{n+1}} - \frac{1}{(1 - e^{-2u})^n} \right\}.$$

The expansion in a series of powers of e^{-u} is quite simple, though the coefficients are somewhat complicated: the result is

$$t^n = 2^n n! \Big\{ J_n + (n+2)J_{n+2} + (n+1)\frac{n+4}{2} J_{n+4} + \cdots + \binom{n+k-1}{k-1} \frac{n+2k}{k} J_{n+2k} + \cdots \Big\}$$

The above results are all well known and are only quoted as an illustration of the procedure; other expansions of known functions in Neumann series can be obtained similarly. The procedure can also be applied to the solution of differential equations in terms of t: the equation should be "Laplace-transformed" in terms of $p = \sinh u$ and the unknown is then multiplied by $\cosh u$; if, by assuming $e^{-u} = z$ as a new independent variable, the solution can be expressed as a power series of z around the origin, the Neumann series having the same coefficients gives a solution of the original equation.

5. It has been observed that when $f(t)$ has some finite singularity, the Λ-transform cannot be expanded in a series of powers of e^{-u}. This means that, in the plane $z = e^{-u}$, the Λ-transform of a function $f(t)$ should be singular in the origin $z = 0$ whenever $f(t)$ has some singularity other than infinity. The circumstance is readily illustrated by an example. A unit step function occurring at the instant $t = a > 0$ has the Laplace transform e^{-ap}/p: the Λ-transform is, accordingly, $\coth u \cdot e^{-a \sinh u}$; and this expression cannot be expanded in series of $z = e^{-u}$, since the exponential term, which can be written $\exp\left[\frac{a}{2}(z - \frac{1}{z})\right]$ has an essential singularity at $z = 0$. By shifting the origin to the discontinuity, the Laplace transform becomes $1/p$, the Λ-transform $\coth u = 1 + 2e^{-2u} + 2e^{-4u} + \cdots$, corresponding to the Neumann series $J_0 + 2J_2 + 2J_4 + \cdots$, whose sum is unity for any value of t. The original function is really unity for $t > 0$ only, but it must be remembered that the transform defines the function in the only region $t > 0$.

It can be seen, from the above example, that the Λ-transform of the discontinuous function takes the form of a regular term, $\coth u$, multiplied by a singular term, $\exp(-a \sinh u)$, where the coefficient a in the exponent is merely the value of t where the discontinuity occurs. The result is easily extended to the case of more than one singularity: it suffices to regard the discontinuous function as resulting from the sum of several continuous functions, having their origin in the successive singularities respectively.

6. As compared with conventional methods the procedure expounded shows all advantages and disadvantages of numerical methods. As is always the case with results expressed in the form of series, it is to be presumed, for the solution to be useful, that the terms beyond a given order may be neglected in the given order of accuracy. The number of terms to be considered for a given behavior of the coefficients, depends on the maximum useful value of the argument of the Bessel functions; and this value depends on the unit in which the independent variable is measured. The choice of a larger unit makes the maximum value of t smaller, but obviously influences the value of the coefficients of the Neumann series, so that a rational choice is a matter of compromise.

When t is small enough, i.e., when we are interested in the initial behavior of a function, the direct numerical computation is immediately possible and requires the determination of a limited number of coefficients only: in such cases, the present procedure can often be preferable to classic methods.

Due to the typical behavior of Bessel functions of the first kind, a Neumann series is a convenient expansion for a function having the character of a damped oscillation—provided the unit is chosen properly. The inversion of a Laplace transform in a Neumann series is thus expected to be useful when it is known that the function sought is of this type.

When the Laplace transform is a rational function of p, the classic method calls for the determination of the zeros of the denominator, that is, for the solution of an algebraic equation. As soon as this step has been performed, the inverse transform is easily expressed as the sum of terms of the form $e^{-at} \sin (bt + \varphi)$ where the damping coefficients, such as a, and the angular velocities, such as b, are determined respectively by the real and imaginary parts of the roots of the denominator. The decomposition of the function into elementary components of the form of damped oscillations describes the function itself completely.

However, the inversion can involve a prohibitive amount of labor, when the degree of the equation is even moderately high. When we are chiefly concerned with the initial behavior of $f(t)$, such as when we are interested in "overshoots", for instance, the problem can be solved numerically by an obvious procedure which remains the same independently of the degree of the denominator. The procedure consists in expanding the given rational function in a series of powers of $1/p$, the coefficients of the series being determined by direct substitution. The rational function of p which represents the Laplace transforms is first expressed as a ratio of polynomials of $1/p$: $f(t) = F(1/p)/G(1/p)$; the expansion in a numerical series of powers of $1/p$ is then possible, by equating term by term the two sides of the equation:

$$F(1/p) = G(1/p) \left\{ \frac{a_1}{p} + \frac{a_2}{p^2} + \frac{a_3}{p^3} + \cdots \right\}. \tag{5}$$

The Λ-transform, accordingly, assumes the forms of a sum of terms $\cosh u / (\sinh u)^n$, that can be individually expanded in power-series of e^{-u}; their

sum gives directly the expansion of the Λ-transform and, consequently, the Neumann series for $f(t)$.

7. The possibility of determining the first coefficients of the Neumann series by simple numerical operations, becomes particularly interesting when the Laplace transform is given in the form of the product of two or more functions of p. This is the case in one of the most useful applications of Laplace transform, that is, when the problem is that of determining the output function of a system whichever, given the input function.

The relation between output and input in a linear system is generally determined by equating two linear differential expressions operating on the input and the output. Symbolically, the output can then be expressed in the form of a rational function of the operational variable $p = d/dt$, multiplied by the input; (in the *steady-state* equations of electrical networks, the variable p reduces to $2\pi j$ times the frequency). This function of p is the "symbolic transfer function" of the system, and gives the actual ratio *between the Laplace transforms* of the output and the input.

Thus, if the problem is that of determining the output, given the input transient and the symbolic transfer function, a natural way of attack is that of Laplace-transforming the input and obtaining the transform of the output by direct multiplication.*

The problem which next arises, that is, that of inverting the transform, is solved classically by the "convolution theorem" which writes in the form of the integral $\int_0^t f(\tau)g(t-\tau)\,d\tau$ the function having as Laplace transform the product of the transforms of f and g.

According to this theorem, the output of the system is expressed as the convolution integral of the input and the function that has the transfer function as Laplace transform. This function is often very complicated, and, accordingly, the computation of the convolution integral may require such an amount of labor as to appear inpractical.

Following the present procedure, the transfer function, expressed as a series of powers of e^{-u} after making $p = \sinh u$†, is multiplied term by term by the series giving the Λ-transform of the input: the result is the Λ-transform of the output. If it has been ascertained that both series can be stopped at a certain term, that is, that they may be replaced by mere polynomials, the multiplication term by term is obviously legitimate, to an accuracy which is easily estimated.

8. As an example of the procedure, let us consider a "line", that is, a unidimensional medium where at any abscissa x and at any time t a cause ("voltage") and an effect ("current") is present.

The current $i(x, t)$ determines in the length dx of the line a voltage drop pro-

* The symbolic transfer function is merely the transform of the response of the system to the "unit pulse".

† It is easily shown that this is the Λ-transform of the response of the system to a function having unity as Λ-transform: i.e., to the Bessel function $J_0(t)$.

portional to i plus a drop proportional to di/dt: $-dv = ri\,dx + l(di/dt)\,dx$; similarly, the voltage determines a current leak of the form $-di = gv\,dx + c(dv/dt)\,dx$. In symbols, writing p for (d/dt), $-(\partial v/\partial x) = (r + lp)i$; $-(\partial i/\partial x) = (g + cp)v$.

The equations, which are symbolic for the functions $v(x, t)$ and $i(x, t)$ become conventional differential equations if v and i are now intended as *the Laplace transforms*, of the voltage and current respectively, with respect to time (under the assumption $v(x, o) = 0$, $i(x, o) = 0$.) The general solution of the equations is

$$y = C_1 e^{\gamma x} + C_2 e^{-\gamma x}$$

where y stands for v or i, $\gamma^2 = (r + lp)(g + cp)$ and C_1, C_2 are arbitrary functions of p. If, for instance, v and i must be finite at infinity, C_1 must be zero, so that

$$v = C_2 e^{-\gamma x}; \qquad i = D_2 e^{-\gamma x}.$$

It is easily found, by substitution, that $D_2/C_2 = \sqrt{(g + cp)/(r + lp)}$, so that the above expressions make it possible to write down the (Laplace transform of) current and voltage at any point of the line, given the transform of one of the variables at one point, e.g. in the origin.

Thus, if $\mathfrak{V}_0$ is the Laplace transform of the voltage at the origin, the transform of the voltage at the abscissa x is

$$\mathfrak{V}_x = \mathfrak{V}_0 \exp\{-x\sqrt{(r + lp)(g + cp)}\} \tag{6}$$

and the function $v(x, t)$ can be obtained as the inverse transform of this expression. To perform the inversion, according to the convolution theorem, we must effect the inversion of the exponential term which can be written as

$$\exp\{-x\sqrt{lc}\,\sqrt{(p + a)^2 - b^2}\}$$

with

$$a = \frac{1}{2}\left(\frac{r}{l} + \frac{g}{c}\right), \qquad b = \frac{1}{2}\left(\frac{r}{l} - \frac{g}{c}\right).$$

The classic theory of Laplace transform writes the inverse transform of the exponential in the form

$$e^{-ax\sqrt{lc}}U(t - x\sqrt{lc}) + bx\sqrt{lc}e^{-at}\frac{I_1(b\sqrt{t^2 - x^2lc})}{\sqrt{t^2 - x^2lc}}$$

where I_1 is the modified Bessel function of the first order and U is the "unit pulse function". For t smaller than $x\sqrt{lc}$ the inverse transform is zero: for $t > x\sqrt{lc}$, $v(x, t)$ is given by the convolution integral of $v(0, t) = v_0(t)$ by the above expression; that is:

$$\begin{aligned} v(x, t) &= e^{-ax\sqrt{lc}}v_0(t - x\sqrt{lc}) \\ &\quad + bx\sqrt{lc}\int_{x\sqrt{lc}}^{t} e^{-a\tau}\frac{I_1(b\sqrt{\tau^2 - x^2lc})}{\sqrt{\tau^2 - x^2lc}}\,v_0(t - \tau)\,d\tau \end{aligned} \tag{7}$$

From the theoretical point of view, the formula solves the problem completely, but the computation can be prohibitively complicated, unless the form of the function $v_0(t)$ is very particular.

Formula (6) can be written in terms of Λ-transforms by simply putting p equal to $\sinh u$ and multiplying both sides by $\cosh u$. The Λ-transform of $v(x, t)$ thus becomes equal to that of $v_0(0, t) = v_0(t)$, multiplied by the exponential, expressed in terms of u. To work with pure numbers, we shall find it convenient to assume $x\sqrt{lc}\cdot p$ (which in dimensionless) as the Laplace variable $\sinh u$: this corresponds to the choice of $x\sqrt{lc}$ as the unit of time. The exponential then becomes.

$$\exp\left\{-\sqrt{\left(\frac{xr}{\rho}+\sinh u\right)(xg\rho+\sinh u)}\right\}$$

$$= \exp\{-\sqrt{(A+\sinh u)(B+\sinh u)}\}$$

where $\rho = \sqrt{l/c}$ is the "characteristic resistance" of the line. To compute the right side of

$$\Lambda\{v(x, t)\} = \Lambda\{v(0, t)\}\ \exp\left\{-\sqrt{\left(\frac{xr}{\rho}+\sinh u\right)(xg\rho+\sinh u)}\right\} \quad (8)$$

where $v(0, t)$ is supposed to be known, and expressed as a power series of e^{-u}, we should have to express the exponential in (8) as a series of the same type. This is not possible, since the exponential is singular at $e^{-u} = 0$, and the circumstance immediately informs of the existence of a singularity. (This is physically evident, since the transient $v(x)$ can only take place after a transmission time, which is just $x\sqrt{lc}$).

It is immediately evident, however, that the singularity in the exponential is of the same nature as that of $e^{-\sinh u}$ so that, if we write

$$\exp\{-\sqrt{(A+\sinh u)\,(B+\sinh u)}\}$$

$$= \exp(-\sinh u)\cdot\exp\{\sinh u - \sqrt{(A+\sinh u)\,(B+\sinh u)}\}$$

the second factor is no longer singular in the origin of e^{-u}, as is easily recognized. The first factor merely shows that the transient takes its beginning at $t = 1$: the second must be expanded in powers of e^{-u}, and this is numerically possible without difficulty. The exponent is first written in the form

$$-F = -\sinh u\left\{\sqrt{\left(1+\frac{A}{\sinh u}\right)\left(1+\frac{B}{\sinh u}\right)}-1\right\} \quad (9)$$

and then expanded in series of powers of $1/\sinh u$. The identity $1/\sinh u = 2e^{-u}/(1 - e^{-2u})$ allows the series to be transformed in terms of e^{-u}.

A shorter procedure consists in writing down a differential equation, in terms of the variable $z = e^{-u}$, satisfied by the function (9), and expanding the latter around the origin.

The equation is found to be

$$z\,[(1 + 2az - z^2)^2 - 4b^2z^2]\,\frac{dF}{dz}$$

$$+ (1 + z^2)(1 + 2az - z^2)\,(F - a) + 2b^2z(1 + z^2) = 0$$

where a and b stand for $(A + B)/2$ and $(A - B)/2$ respectively. The initial value is $F = a$; and the series around the origin becomes of the form

$$F = a + g_1z + g_2z^2 + \cdots \tag{10}$$

with $g_1 = -b^2$, $g_2 = 2ab^2$, and the successive coefficients given by a rather complicated recurrence formula. We note incidentally that g_3 is of the type

$$-b^2 + [\text{terms of the order of } b^4].$$

As has been repeatedly stated, the interest of the method expounded is chiefly in the cases where a small number of terms suffice, and this is the case in normal lines where the constants a and b are actually small (in *distortionless* lines b is zero). In such cases, neglecting terms of the order of b^3, we have

$$F \cong a - b^2z(1 + z^2) \tag{11}$$

In any case, given the series (10) for F, a numerical series of the type

$$e^{-F} = e^{-a}\{1 + h_1z + h_2z^2 + \cdots\} \tag{12}$$

is easily obtained, at least if a limited number of terms suffice. In the approximation of formula (11) we can write

$$e^{-F} = e^{-a}[1 + b^2z(1 + z^2)]^*$$

Once the series (12) has been constructed, formula (8) becomes

$$\Lambda\{v(x, t)\} = \Lambda\{v(0, t)\} \cdot e^{-\sinh u} \cdot e^{-a}[1 + h_1z + h_2z^2 + \cdots]$$

where $\Lambda\{v(0, t)\}$ is supposed to be expansible in a power series of z. The multiplication term by term gives the expansion of the left side; the factor $e^{-\sinh u}$ indicates that the transient begins at $t = 1$; the factor e^{-a} is merely a constant.

Bibliography

The argument of Laplace transform is discussed in most books of Analysis. A classic treatise is that of G. Doetsch, *Theorie und Anwendung der Laplace Transformation*, Springer, Berlin, 1937; American reprint 1943, Dover Publications, New York. A shorter (but, for the present purpose, sufficient) exposition is contained in Chapter 1 of vol. 18 of the well-known Radiation Laboratory Series of M.I.T. (Vacuum Tube Amplifiers).

Formulas concerning Bessel Functions are found in any handbook of higher

* It will be of interest to remark, as an example, that in an electrical trasmission line of normal characteristics, b is of the order of one per cent per kilometer, so that b^3 is of the order of one part in a million, on the length of one kilometer.

mathematics. A complete treatise is that of G. N. Watson, *Bessel Functions*, Cambridge 1944 (2nd edition); also A. Gray, G. B. Mathews and T. M. MacRobert, *Bessel Functions*, London 1931; N. W. McLachlan, *Bessel Functions for Engineers*, Oxford, 1941 (2nd edition).

Neumann Series form the topic of Chapter XVI (p. 522–550) of Watson's treatise. Many interesting results, that could be utilized for further development of the procedure expounded above, can be found in this Chapter. It will be noted, for instance, that the procedure is substantially equivalent to Cailler's result on Neumann series (p. 536). Cailler showed that, if $f(t) = \sum a_n J_n(t)$, then

$$\sum a_n z^n = \frac{1+z^2}{1-z^2} \int_0^\infty e^{-t} f\left[\frac{2zt}{1-z^2}\right] dt$$

Making $z = e^{-\alpha}$ the integral becomes equivalent to (2).

The computation of a function given in the form of a Neumann series can often be radically shortened by the knowledge of some series of known sum, such as those recalled in paragraph 4, that can be subtracted from the series obtained; in some cases this allows the sum of the series to be written in closed form.

If this is not the case, the computation is only possible if a table giving all significant J_n is available. The 12-volume table of the Harvard Computation Laboratory (1947–1951) is the most complete table of this kind. J_n is given for $n = 0$ to 135, $x = 0$ to 100: J_1 to J_3 are given to 18 decimal places, higher n's to 10 places. The interval in x is 0.001 for the smaller values of x and n, 0.01 elsewhere.

The table of the British Association for the Advancement of Science (Vol. VI and X of the Mathematical Tables, Cambridge 1937–1952) is also recommended. In addition to J_n, (8 or 10 decimal places, n from 0 to 20, x from 0 to 25, interval 0.01 or 0.1), the table also gives the other Bessel Functions (Y_n, I_n, K_n).

The smaller table of the author (Eleven- and Fifteen-place Tables of Bessel Functions, New York, 1948) was just planned in view of its application to the computation of Neumann series. Thus, for every value of x, all "significant" J_n (i.e. all values greater than one half unit of the last place) are given. The interval is 0.001 for x smaller than 0.5 (15 places), 0.01 up to $x = 10.5$ (11 places).

Acknowledgement. The author wishes to express his indebtedness to Prof. Vittorio Gori, President of the Foundation "Ugo Bordoni", for advice and encouragement.

ROMA, FONDAZIONE "UGO BORDONI"

(Received April 12, 1954)

A METHOD OF SOLVING INHOMOGENEOUS LINEAR SIMULTANEOUS EQUATIONS

By F. B. Stulen and F. G. Lehman

Introduction. An iteration method is presented here for the solution of a set of linear simultaneous equations. The development is limited to the inhomogeneous type of equations with emphasis on those systems which possess real characteristic roots. The method is particularly suitable to problems arising in physics and engineering (i.e., where the characteristic values have physical meaning). In fact, this procedure has been in use for the last eight years in the practical solution of propeller blade static and dynamic problems.

In previous papers by Kincaid (1) and Flanders and Shortley (2) the polynomial operator was used to obtain the characteristic roots and vectors of matrices and characteristic value equations. The following development will show the application of this operator to the inhomogeneous problem.

Development. Assume the set of equations to be solved to be written in the following form for the purpose of solving by iteration:

$$y = b + Ay \tag{1}$$

where y and b are column matrices and A is a square matrix. A convergent solution for y is required. Associated with equation (1) is the homogeneous characteristic value equation

$$\lambda y = Ay \tag{1a}$$

Let the initial assumption for y be

$$y_0 = y + \sum_i a_i Y_i \tag{2}$$

where Y_i are the characteristic functions of A and λ_i are corresponding characteristic values which are the solutions of equation (1a).

Let the iteration process be defined as $y^{(p)} = b + Ay^{(p-1)}$ where $y^{(p)}$ symbolizes the p^{th} iterate of y.

Then substituting (2) in (1), iterating p times, and noting from (1a) that $\lambda_i Y_i = AY_i$ the following is obtained:

$$y^{(p)} = y + \sum_i a_i \lambda_i^p Y_i \tag{3}$$

Thus it is seen that y_p converges to y in the limit when all $|\lambda_i| < 1$.

It is now desired to alter the method of iteration in order to effectively make $|\lambda_i| < 1$. Multiply each of the first m iterates by a coefficient C_k and add.

$$y^{(m1)} = (C_1 + C_2 + \cdots + C_m)y + \sum_i (C_1\lambda_i + C_2\lambda_i^2 + \cdots + C_m\lambda_i^m)a_i Y_i$$

Let the coefficients be subject to the condition $\sum_j C_j = 1$. Then

$$y^{(mq)} = y + \sum_i (C_1\lambda_i + C_2\lambda_i^2 + \cdots + C_m\lambda_i^m)^q a_i Y_i \tag{4}$$

off errors. Putting these origin shifts in factor form the following polynomial is obtained:

$$P(\lambda) = (\lambda - \lambda_1)^{\alpha_1}(\lambda - \lambda_2)^{\alpha_2} \cdots (\lambda - \lambda_{n-1})^{\alpha_{n-1}}(\lambda - \lambda_n)^{\alpha_n} \tag{5}$$

where the α_i are small integers corresponding to the order of moments about the particular origin, λ_i. It is only required that $| P(\lambda) | < 1$ for all λ_i but for rapid convergence $| P(\lambda) | \ll 1$ for all λ_i. The dotted curve shown in Figure 1 would meet this latter requirement in a practical way.

Mechanics of the Method. In order to obtain the proper polynomial it is necessary to first place the nodes in the proper position and second to be able to adjust the amplitudes of the loops. The first requires knowledge of the approximate location of generally the first several characteristic values of the matrix A. Aitken (3) has given a method of estimating the k greatest γ_i which are obtained from the iterates of the characteristic value problem. Frequently estimates of the first several characteristic values of a problem can be made after having some experience with similar problems. Another method of obtaining the k highest characteristic values is to use the convergence polynomial. The dominant mode is found by ordinary iteration and the mode at the other extremity of the range is found by ordinary iteration with the origin at the dominant mode. The subdominant mode is found by passing the convergence polynomial through the extremities of the range, using a larger value for the α associated with the smallest characteristic value. Generally that is as far as one has to go but if necessary the process can be continued. The adjustment in the relative amplitude of the loops is made by choosing the proper value of α_i in equation (5). Smaller amplitudes are obtained by using higher values of α_i. This would be particularly desirable when there are characteristic values between the modes. Each factor in (5) represents an origin shift and the exponent, α_i is the number of iterations performed with that particular origin location in case there are characteristic values located in between the various origins to be used.

In order to shift the origin to the right by the amount λ_i it is necessary to subtract $\lambda_i y$ from both sides of equation (1) resulting in:

$$(I - \lambda_i)y = b + (A - \lambda_i)y$$

which yields

$$y = [b + (A - \lambda_i)y]/(1 - \lambda_i) \tag{6}$$

After completing m origin shifts and $\alpha_1 + \alpha_2 + \cdots + \alpha_k$ iterations one iteration cycle is completed. The second cycle is a repetition of the first one using the values obtained at the end of the first cycle as the initial y_{m1}.

The following example problem using this method of iteration shows that it is sometimes necessary to find fairly accurate characteristic values in order to obtain practical convergence. However, in certain classes of design problems

Equation (2c) will converge to y when the polynomial,

$$|(C_1\lambda_i + C_2\lambda_i^2 + \cdots + C_m\lambda_i^m)| < 1$$

for all i. This type of polynomial will hereafter be referred to as the convergence polynomial or polynomial operator.

Geometrical Interpretation. After studying the physical picture of the characteristic value spectrum, the method of choosing an efficient convergence polynomial will become clearer.

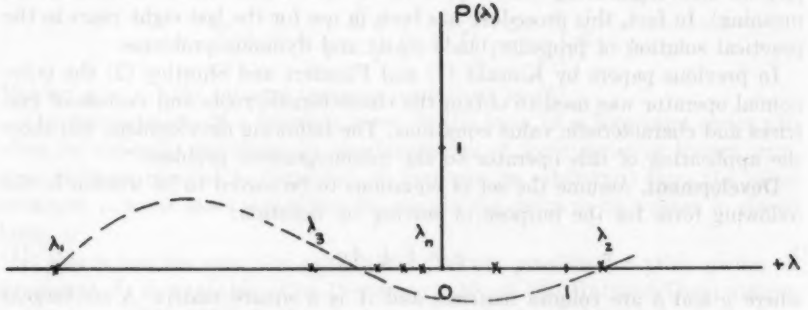


FIG. 1. Characteristic value distribution and polynomial $P(\lambda)$

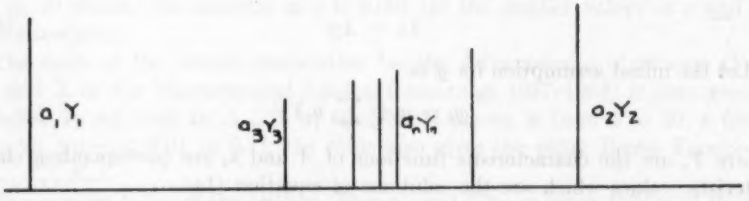


FIG. 2. Distribution of characteristic functions in y_0

Assume a distribution of characteristic values as shown in Figure (1) and a corresponding distribution of the characteristic functions which are the components of y_0 under the summation in (2). Following out the iteration process graphically from Figures (1) and (2) as prescribed by equation (3), it is seen that the p^{th} iterate except for the column y is equivalent to taking the p^{th} order moment of the ordinates of Figure (2) about the origin or about the point $\lambda = 0$. Therefore, the components corresponding to the large values of λ are greatly magnified by successive iterations even though these components themselves might be chosen relatively small.

By shifting the origin of moments to the position of each λ_i the moment components of all characteristic functions are all made zero except for round

APPROXIMATION AND PROJECTION

By Arthur Sard[1]

1. Introduction. Suppose that K is a given operator, $\mathfrak{Z}$ is a given class of operators, and $f = f(x)$ and $\delta f = \delta f(x)$ are functions of x, δf being the error in $f + \delta f$. We shall approximate Kf by $T(f + \delta f)$, where $T \varepsilon \mathfrak{Z}$. In view of the anticipated instances of f and δf, what element T of $\mathfrak{Z}$ should we use?

We may think of constructing a machine to perform T. The class $\mathfrak{Z}$ may consist of those operators that can be mechanized in an acceptable way. The input of the machine will be the contaminated signal $f + \delta f$; the output will be $T(f + \delta f)$. The output is to approximate Kf, where f is the ideal signal. Thus T is to be chosen so as to be in $\mathfrak{Z}$, to approximate K, and to filter out the effect of δf.

As an illustration, consider a speedometer: $f + \delta f$ may be the angular distance travelled at time x by the head of the speedometer cable; $T(f + \delta f)$ the speedometer reading at time x; δf the contribution of error to the angular travel of the head of the cable; $Kf = df/dx$ the true speed at time x; $\mathfrak{Z}$ the class of operators that can be mechanized as speedometers.

Alternatively we may think of T as a process to be executed by a mathematician who wishes to approximate Kf by $T(f + \delta f)$ in a number of instances involving different data f and δf, but one K and one T. Here $\mathfrak{Z}$ may consist of all operators which for one reason or other are acceptable to the mathematician.

In many cases the error δf is due to experiment or to rounding or to approximations of calculation. A table of "values of a function f", for example, is almost always contaminated by errors. What is tabulated is $f + \delta f$, not f. Any routine process for approximating Kf on the basis of a table of values of f is an instance of the use of an operator T so that $T(f + \delta f)$ approximate Kf. In particular K may be the identity or an integration or a q-fold differentiation. Also, δf may vanish identically.

In each approximation the mathematician or calculator deals with one function f and one error δf; and f and δf are often determined by an outside source. We shall therefore suppose that f and δf are stochastic processes: chance produces each. The nature of the stochastic processes will enter in the determination of which element T of $\mathfrak{Z}$ to use.

The exact formulation of the problem will affect our entire analysis. (That formulation will be given in later sections.) The rationale of our specific assumptions is the following. Suppose that $f + \delta f \varepsilon F_1$, $Kf \varepsilon F_2$, where F_1, F_2 are topological spaces; and that each element of $\mathfrak{Z}$ is an operator on F_1 to F_2. We shall assume that the operators in $\mathfrak{Z}$ are continuous. This is reasonable, since otherwise a small change in the argument of an operator might induce a large change in the approximation. Continuity however depends on the topologies of

[1] Research sponsored by the Office of Ordnance Research, U. S. Army, under contract DA-30-069-ORD-956.

the distribution of the characteristic values remains nearly constant and it is only necessary to obtain the range of these values. In this example, the exact characteristic values are -1, 0.5, 2 and 4. The origins used for one iteration cycle were 4.001, -1.001, .4, .4, 2.1 and 2.1. The availability of high speed digital computing equipment is assumed for practical application of the method.

Example Problem.

System of Equations					Solution @ Cycle			
					0	2	4	Exact
$y_1 = -1$	$-11y_1$	$-8y_2$	$+30y_3$	$+53y_4$	1	65.92	64.84	64.83
$y_2 = -2$	$+4y_1$		$-12y_3$	$-17y_4$	1	97.70	104.64	104.66
$y_3 = -4$	$-11y_1$	$-4y_2$	$+32y_3$	$+49y_4$	1	-99.93	-109.80	-109.83
$y_4 = -8$	$+4y_1$	$+0.5y_2$	$-12y_3$	$-16.5y_4$	1	86.25	92.64	92.66

REFERENCES

(1) KINCAID, W. M., *Numerical Methods for Finding Characteristic Roots and Vectors of Matrices*, Quarterly of Applied Mathematics, Vol. 5, No. 3, 1947.

(2) FLANDERS, D. A. AND SHORTLEY, G., *Numerical Determination of Fundamental Modes*, Journal of Applied Physics, Vol. 21, December, 1950.

(3) AITKEN, A. C., *Studies in Practical Mathematics, II The Evaluation of the Latent Roots and Latent Vectors of a Matrix*, Proceedings of the Royal Society of Edinburgh, Vol. LVII, Part III, 1936–37.

CURTISS-WRIGHT CORPORATION,
PROPELLER DIVISION

(Received May 20, 1955)

the spaces F_1 and F_2. If a mathematician is interested in a class $\mathfrak{Z}$ of operators which are not continuous, he may be able to find new topologies which are suitable for his purpose and relative to which the operators of $\mathfrak{Z}$ are continuous. For example in some cases the size of $f(x)$ will depend on an integral of $| f(x) |$; in other cases on an integral of $| f'(x) |$; and so on in similar and dissimilar ways[2].

Besides requiring that the elements of $\mathfrak{Z}$ be continuous on F_1 to F_2, we shall further limit our consideration as follows. We shall assume that F_1, F_2 are Hilbert spaces. This assumption is relatively light since so many different function spaces are Hilbert spaces. We shall assume that the operators in $\mathfrak{Z}$ are linear. This restriction is severe but fruitful. So many actual approximations are indeed linear and continuous on properly chosen Hilbert spaces that our problem has broad applicability. We shall assume that $\mathfrak{Z}$ is a linear set. This assumption is relatively light since we may, if necessary, replace $\mathfrak{Z}$ at the beginning of our analysis by the linear span of $\mathfrak{Z}$.

We now give an indication of our results. Let

$$e = T(f + \delta f) - Kf$$

be the error in the approximation of Kf by $T(f + \delta f)$; suppose that e is a function of a variable y. We say that T in $\mathfrak{Z}$ is nearly efficient if $E \| e \|^2$ is near its infimum over $\mathfrak{Z}$, where E indicates expected value and $\| e \|^2$ is the integral of $| e |^2$ relative to a given measure on the space Y of the argument y; and we say that T in $\mathfrak{Z}$ is nearly strongly efficient if, essentially, $E \mid e \mid^2$ is near its infimum for each y in Y (§3). Nearly efficient operators always exist. An operator which is nearly strongly efficient is surely nearly efficient, but not vice versa.

A necessary and sufficient condition that an operator T_1 in $\mathfrak{Z}$ be nearly efficient is that $T_1(f + \delta f)$ be near the projection of Kf on the subspace L which is the closure of all elements $T(f + \delta f)$, $T \in \mathfrak{Z}$ (Theorem 1). Necessary and sufficient conditions for near strong efficiency are given in Theorems 3, 4.

We say that an operator T is of finite rank if the image under T of the space on which T operates is finite dimensional. If $\mathfrak{Z}$ contains all operators of finite rank, efficiency is necessarily strong (Theorem 5). Thus if $\mathfrak{Z}$ is sufficiently large,

[2] To approximate Kf in terms of $f + \delta f$, one might use conditional expectations. Let

$$\bar{T}(f + \delta f) = E(Kf \mid f + \delta f),$$

that is, $\bar{T}(f + \delta f)$ is the conditional expectation of Kf, given $f + \delta f$ [2, pp. 77, 78]. The operator $\bar{T}$, in certain cases at least, is well defined and minimizes $E \mid T(f + \delta f) - Kf \mid^2$ for each value of the argument of Kf among *all* operators T. The operator $\bar{T}$ is unbiased: $E\bar{T}(f + \delta f) = Kf$. Nonetheless we shall not consider $\bar{T}$ in this paper because, even in simple cases, $\bar{T}$ may be discontinuous in all natural topologies. To see this, consider the following example. Let F_1, F_2 each be the space of complex numbers, so that f and δf are merely random variables. Let probability be Lebesgue measure on $-.5 \leqq \omega < .5$. Let $f = 2\pi\omega$, $\delta f = \sin 2\pi\omega - 2\pi\omega$, $K =$ the identity. Then $Kf = 2\pi\omega$, $f + \delta f = \sin 2\pi\omega$, and $\bar{T}(f + \delta f) = E(2\pi\omega \mid \sin 2\pi\omega)$ has three values: $\bar{T}(f + \delta f)$ equals $\pi/2$ if $\sin 2\pi\omega > 0$; 0 if $\sin 2\pi\omega = 0$; $-\pi/2$ if $\sin 2\pi\omega < 0$. Thus $\bar{T}(\sin 2\pi\omega)$ jumps as $\sin 2\pi\omega$ passes thru 0, and $\bar{T}z$ is discontinuous if the norm of z is $| z |$. There is no alternative norm which is natural and which makes $\bar{T}$ continuous.

an operator which is nearly efficient is also nearly strongly efficient. This result seems interesting and important.

An answer to the question as to which T in $\mathfrak{T}$ should be used in the approximation of Kf by $T(f + \delta f)$ is the following: T should be nearly efficient. If T is then nearly strongly efficient, so much the better.

The idea of least square approximation as a projection in an appropriate space appears in Nikodym [8], Weyl [12], Kolmogorov [5], Doob [2; pp. 78, 155], and elsewhere. Our formulation of the problem of this paper (§§2, 3) in terms of direct products of Hilbert spaces seems apt and new as far as the theory of approximation is concerned. Kolmogorov [4] and Wiener [13] have studied what we call strongly efficient operators in the case in which f and δf are stationary stochastic processes with zero means [2; Chap. 12]. Wiener's prediction operator is our strongly efficient operator in the following particular instance:

x is time; $-\infty < x < \infty$;

f and δf are functions of x;

K is the operation of lead by $c \geqq 0$: $K[\varphi(x)] = \varphi(x + c)$;

$\mathfrak{T}$ is the class of operators on the past, for example, all operators T of the form

$$T\varphi = \int_0^\infty \varphi(x - \xi) d\tau(\xi).$$

An efficient operator is similar to but not necessarily the same as the approximation process of our earlier paper [10]. We note in passing that the first part of [10] (up to and including Theorem 1 but excluding the Corollary) can be generalized to include the approximation of Kf rather than of merely f.

2. The Hilbert spaces. Let Y be a space and m a measure, not identically zero, on Y. Let H_Y be the space of functions absolute square integrable m, that is, the space of complex functions $u = u(y)$, $y \in Y$, measurable m such that

$$\| u \|_Y^2 = \int_Y | u(y) |^2 \, dm(y) < \infty,$$

it being understood that two functions which are equal almost everywhere m correspond to the same element of H_Y and that the inner product of u into v is

$$(u, v)_Y = \int_Y u(y)\bar{v}(y) \, dm(y), \qquad u, v \in H_Y.$$

We assume that the measure m is such that H_Y is a Hilbert[3] space [3, pp. 177, 168; 9, II p. 42 and I pp. 104, 170].

Let Ω be a probability space, that is, a space with probability p and expected value operator E:

$$Ez = \int_\Omega z \, dp, \qquad E1 = 1.$$

[3] By a *Hilbert space* we mean a separable, complete, linear space over the complex numbers with an inner product, whether the dimension of the space be finite or countably infinite. We assume that the dimension is not zero.

Let H_Ω be the space of functions $u = u(\omega)$, $\omega \varepsilon \Omega$, absolute square integrable p. We assume that p is such that H_Ω is a Hilbert space. Note that

$$1 \varepsilon H_\Omega, \qquad (u, v)_\Omega = Eu\bar{v}, \qquad (u, 1)_\Omega = Eu, \qquad \| u \|_\Omega^2 = E \mid u \mid^2,$$

where $u, v \varepsilon H_\Omega$.

Let $H_{Y\Omega}$ be the space of functions $u = u(y, \omega)$, $y \varepsilon Y$, $\omega \varepsilon \Omega$, absolute square integrable mp. Then $H_{Y\Omega}$ is a Hilbert space, and the inner product in $H_{Y\Omega}$ is

$$(u, v)_{Y\Omega} = E(u, v)_Y = E \int_Y u\bar{v}\, dm, \qquad u, v \varepsilon H_{Y\Omega}.$$

When $u \varepsilon H_{Y\Omega}$, we will write variously

$$u = u(y, \omega) = u(y) = u(\omega) = u_y = u_\omega,$$

depending on considerations of emphasis and typography.

Note that $H_{Y\Omega}$ is the direct product $H_Y \times H_\Omega$ of H_Y into H_Ω; that is, $H_{Y\Omega}$ is the closure of the span[4] of the elements

$$\psi_{j,\nu} = \psi_{j,\nu}(y, \omega) = u_j(y)\varphi_\nu(\omega), \qquad j \varepsilon J, \nu \varepsilon N,$$

the inner product being determined by the relations

$$(\psi_{j,\nu}, \psi_{j',\nu'})_{Y\Omega} = \delta_{j,j'}\delta_{\nu,\nu'},$$

where $u_j(y)$, $j \varepsilon J$, is any complete orthonormal set in H_Y and $\varphi_\nu(\omega)$, $\nu \varepsilon N$, is any complete orthonormal set in H_Ω [7, pp. 127–133, particularly Lemma 2.2.1; 11].

Let H_X be a Hilbert space, perhaps like the space H_Y, but not necessarily so. Let $H_{X\Omega}$ be the direct product $H_X \times H_\Omega$.

The following lemmas state certain natural facts about the above spaces. We omit the proofs. The suffix X may be replaced by Y, but Y not necessarily by X.

Lemma 1. *If $u \varepsilon H_{X\Omega}$, then for each fixed ω with null[5] exceptions $u_\omega \varepsilon H_X$; also $Eu \varepsilon H_X$. If $u \varepsilon H_{Y\Omega}$, then for each fixed y with null exceptions $u_y \varepsilon H_\Omega$.*

Lemma 2. *Suppose that the sequence $\{u^\nu\}$, $\nu = 1, 2, \cdots$, $u^\nu \varepsilon H_{Y\Omega}$, converges in $H_{Y\Omega}$ to u. Then $\{Eu^\nu\}$ converges in H_Y to Eu. For a properly chosen subsequence $\{u^{\nu_n}\}$, $n = 1, 2, \cdots$, $\{u^{\nu_n}(\omega)\}$ converges in H_Y to $u(\omega)$ for almost all ω and $\{u^{\nu_n}(y)\}$ converges in H_Ω to $u(y)$ for almost all y.*

Lemma 3. *Suppose that T is a bounded linear operator on H_X to H_Y. There exists a unique bounded linear operator T' on $H_{X\Omega}$ to $H_{Y\Omega}$ such that $T'(\varphi v) = \varphi T(v)$ whenever $v \varepsilon H_X$, $\varphi \varepsilon H_\Omega$. In particular φ may be unity. The Banach norm of T' equals that of T. If $u \varepsilon H_{X\Omega}$, then $TEu = ET'u$; also for each fixed ω with null exceptions*

$$T(u_\omega) = (T'u)_\omega.$$

[4] By the span of a set A, we mean the set of all finite linear combinations with complex coefficients of elements of A.

[5] Phrases like "null", "measure zero", "almost everywhere" refer to that measure which is appropriate to the context. For example, "each ω with null exceptions" means "each ω in Ω_1, where $p(\Omega - \Omega_1) = 0$"; "almost all y" means "all y in Y_1, where $m(Y - Y_1) = 0$".

We consider T' here as a unique extension of T, and we shall henceforth write T instead of T', since the context will indicate whether T is to be considered as an operator on H_X to H_Y or on $H_{X\Omega}$ to $H_{Y\Omega}$.

3. Efficient and strongly efficient operators. *Let there be given an element g of $H_{X\Omega}$, an element h of $H_{Y\Omega}$, and a linear set $\mathfrak{Z}$ of bounded linear operators on H_X to H_Y. We shall approximate h by Tg, where $T \varepsilon \mathfrak{Z}$.* The present paper is concerned with the study of which elements T in $\mathfrak{Z}$ are best for the approximation.

For each ω with null exceptions, g_ω is an element of H_X, and $(Tg)_\omega = Tg_\omega$ and h_ω are elements of H_Y (Lemmas 1, 3). The element h_ω is approximated by the transform of g_ω under T, where $T \varepsilon \mathfrak{Z}$. The elements g_ω of H_X and h_ω of H_Y may be thought of as determined stochastically. The probability p which entered in the definition of H_Ω describes the anticipated relative importance of the different inputs g_ω and desired outputs h_ω.

The present approximation becomes that of the introduction if we put $g = f + \delta f$ and $h = Kf$. We do not require, however, that h be of the form Kf.

In considering operators T_1 and T, we put

$$e_1 = T_1 g - h, \qquad e = Tg - h.$$

We say that T_1 in $\mathfrak{Z}$ is *efficient* if

$$E \| e_1 \|_Y^2 \leqq E \| e \|_Y^2 \tag{1}$$

whenever $T \varepsilon \mathfrak{Z}$; and we say that T_1 in $\mathfrak{Z}$ is *strongly efficient* if

$$E \mid e_1(y) \mid^2 \leqq E \mid e(y) \mid^2 \qquad \text{for almost all } y \tag{2}$$

whenever $T \varepsilon \mathfrak{Z}$. (The null set on which the inequality does not hold may depend on T.)

Let $\epsilon > 0$ be given. We say that T_1 in $\mathfrak{Z}$ is *within ϵ of efficiency* if

$$E \| e_1 \|_Y^2 < E \| e \|_Y^2 + \epsilon^2 \tag{3}$$

whenever $T \varepsilon \mathfrak{Z}$; and we say that T_1 in $\mathfrak{Z}$ is *within ϵ of strong efficiency* if a function $\zeta = \zeta(y)$ exists such that

$$\zeta \varepsilon H_Y, \qquad \| \zeta \|_Y < \epsilon \tag{4}$$

and such that

$$E \mid e_1(y) \mid^2 \leqq E \mid e(y) \mid^2 + \mid \zeta(y) \mid^2 \qquad \text{for almost all } y \tag{5}$$

whenever $T \varepsilon \mathfrak{Z}$.

Remark. An *unbiased* operator T is one for which $Ee = 0$. One may wish in the above definitions to consider only those elements T in $\mathfrak{Z}$ which are unbiased. In other words one may wish to minimize $E \| e \|_Y^2$ and $E \mid e(y) \mid^2$ among all unbiased T in $\mathfrak{Z}$ rather than among all T in $\mathfrak{Z}$. Such minimizations may be reduced to the ones just formulated in the following way. Let T_0 be an arbitrary, fixed, unbiased element of $\mathfrak{Z}$. Let $\mathfrak{S}$ be the subset of $\mathfrak{Z}$ consisting of all elements S of $\mathfrak{Z}$

such that $ESg = SEg = 0$ (Lemma 3). The set $\mathfrak{S}$ is linear. The set of all unbiased elements of $\mathfrak{Z}$ is $T_0 + \mathfrak{S}$, that is, all elements of the form $T_0 + S$, $S \in \mathfrak{S}$. Suppose that $T = T_0 + S$. Then

$$e = Tg - h = (T_0 + S)g - h = Sg - (h - T_0 g).$$

Thus the error in the approximation of h by Tg is precisely the error in the approximation of $h - T_0 g$ by Sg. Accordingly to add the restriction of unbiasedness to our above definitions of efficiency and strong efficiency, it is sufficient merely to replace $\mathfrak{Z}$ by $\mathfrak{S}$, and h by $h - T_0 g$.

4. Characterization of efficient operators. *Let*

$$L^0 = \mathfrak{Z}g$$

be the set of all elements Tg, $T \in \mathfrak{Z}$. Since $\mathfrak{Z}$ is linear, L^0 is a linear subspace of $H_{r\Omega}$. *Let* L *be the closure of* L^0.

If L^0 is finite dimensional, then L^0 is surely closed and hence identical with L. Let

$$\theta = \mathrm{Proj}_L\, h;$$

that is, θ is the orthogonal projection in $H_{r\Omega}$ of h on L. Then surely $\theta \in L$.

Theorem 1. *An efficient operator exists if and only if* $\theta \in L^0$. *Furthermore* T *in* $\mathfrak{Z}$ *is efficient if and only if*

$$Tg = \theta. \tag{6}$$

Given $\epsilon > 0$. *An operator within* ϵ *of efficiency surely exists. Also* T *in* $\mathfrak{Z}$ *is within* ϵ *of efficiency if and only if*

$$E \| Tg - \theta \|_r^2 < \epsilon^2. \tag{7}$$

Proof. The theorem is an immediate consequence of the theorem of Pythogaras. Thus put

$$\eta = \theta - h.$$

Since η is of the form $u - \mathrm{Proj}_L\, u$,

$$\eta \perp L. \tag{8}$$

Consider any T in $\mathfrak{Z}$ and put $e = Tg - h$. Then

$$e - \eta = Tg - \theta \in L. \tag{9}$$

Hence, by (8), η and $e - \eta$ lie in orthogonal complements in $H_{r\Omega}$. Hence

$$\begin{aligned} \| e \|_{r\Omega}^2 &= \| \eta \|_{r\Omega}^2 + \| e - \eta \|_{r\Omega}^2, \\ E \| e \|_r^2 &= E \| \eta \|_r^2 + E \| e - \eta \|_r^2, \end{aligned} \tag{10}$$

since $e = \eta + (e - \eta)$.

Given $\epsilon > 0$. There is a T in $\mathfrak{Z}$ such that

$$E \| e - \eta \|_r^2 = E \| Tg - \theta \|_r^2 < \epsilon^2, \tag{11}$$

since θ is a limit point of $L^0 = \mathfrak{Z}g$. Hence (10) implies that

$$\inf_{T\varepsilon\mathfrak{Z}} E \parallel e \parallel_Y^2 = E \parallel \eta \parallel_Y^2 ;$$

also that T in $\mathfrak{Z}$ is within ϵ of efficiency if and only if T satisfies (11) or (7).

Also T in $\mathfrak{Z}$ is efficient if and only if

$$E \parallel e - \eta \parallel_Y^2 = E \parallel Tg - \theta \parallel_Y^2 = 0.$$

This establishes (6).

Finally if $\theta \varepsilon L^0$, then $\theta = Tg$, $T \varepsilon \mathfrak{Z}$ and T is efficient. Conversely if an efficient operator exists, $\theta \varepsilon L^0$ by (6).

Corollary. *An operator T in $\mathfrak{Z}$ is efficient if and only if*

$$e = Tg - h \perp L.$$

Proof. If T is efficient, $e = \eta \perp L$ by (6), (8).

Conversely if $e \perp L$, then $e - \eta \perp L$ by (8). Also $e - \eta \varepsilon L$ by (9). Hence $e - \eta = 0$.

Remark. We may describe θ as the efficient approximation and η as the error in efficient approximation. These names are apt because of the above theorem. Explicit formulas for θ and η may be obtained as follows. Let $\psi^\lambda, \lambda \varepsilon \Lambda$, be a complete orthonormal set in L and let $\psi_\mu, \mu \varepsilon M$, be a complete orthonormal set in $L^\perp$. Then, in the sense of convergence in the mean square relative to mp (that is, in the sense of the norm in $H_{Y\Omega}$),

$$\theta = \theta(y, \omega) = \sum_{\lambda\varepsilon\Lambda} \theta_\lambda \psi^\lambda = \sum_{\lambda\varepsilon\Lambda} \theta_\lambda \psi^\lambda(y, \omega),$$

$$\eta = \eta(y, \omega) = \sum_{\mu\varepsilon M} \eta_\mu \psi_\mu = \sum_{\mu\varepsilon M} \eta_\mu \psi_\mu(y, \omega),$$

where

$$\theta_\lambda = (h, \psi^\lambda)_{Y\Omega} = (\theta, \psi^\lambda)_{Y\Omega}, \qquad \lambda \varepsilon \Lambda,$$

$$\eta_\mu = -(h, \psi_\mu)_{Y\Omega} = (\eta, \psi_\mu)_{Y\Omega}, \qquad \mu \varepsilon M.$$

5. Can θ be adjoined to L^0? Sections 5, 6, 7, although interesting, are not needed for other parts of this paper.

Suppose that θ is not an element of L^0. We consider the possibility of enlarging $\mathfrak{Z}$ so that the enlarged L^0 include θ.

Since $\theta \varepsilon L$, θ is a limit point of L^0. Hence a sequence $\{U^\nu\}$, $\nu = 1, 2, \cdots$, exists such that

$$U^\nu g \to \theta, \qquad U^\nu \varepsilon \mathfrak{Z}.$$

Then by Lemmas 1, 2, 3 a subsequence $\{T^\nu\}$ of $\{U^\nu\}$ and a set Ω_1 which is almost all of Ω exist such that for all ω in Ω_1,

$$g_\omega \varepsilon H_X, \qquad \theta_\omega \varepsilon H_Y ;$$

$$(T^\nu g)_\omega = T^\nu g_\omega, \qquad \nu = 1, 2, \cdots ; \tag{12}$$

$$(T^\nu g)_\omega \to \theta_\omega .$$

Let G^0 be the span in H_X of g_ω for all ω in Ω_1; and let G be the closure of G^0.

Since $\theta \in L^0$, it is natural to consider the possibility of adjoining to $\mathfrak{Z}$ an operator U on H_X to H_Y such that $Ug = \theta$.

Lemma 4. *There is a unique linear operator U on G^0 to H_Y such that*

$$Ug_\omega = \theta_\omega, \qquad \omega \in \Omega_1. \tag{13}$$

Proof. If U is linear, the relation (13) implies that

$$U \sum_j c_j g_{\omega_j} = \sum_j c_j \theta_{\omega_j} \tag{14}$$

whenever $\omega_j \in \Omega_1$, c_j is a complex number, and j varies over finite sets. Now (14) unambiguously defines a linear operator U on G^0 to H_Y providing that $U0 = 0$, that is, providing that the second member of (14) vanishes whenever the argument of U in the first member vanishes. But whenever that argument vanishes,

$$\sum_j c_j \theta_{\omega_j} = \sum_j c_j \lim T' g_{\omega_j} = \lim T' \sum_j c_j g_{\omega_j} = \lim 0 = 0$$

by (12). Thus the lemma is proved.

The operator U of the above lemma may or may not be bounded on G^0. In order that U be bounded it is necessary and sufficient that $B < \infty$, where

$$B = \sup \frac{\| \sum_j c_j \theta_{\omega_j} \|_Y}{\| \sum_j c_j g_{\omega_j} \|_X}$$

and the supremum is taken as follows: $\omega_j \in \Omega_1$, c_j is a complex number, j varies over finite sets, and $\| \sum_j c_j g_{\omega_j} \|_X \neq 0$. If $B < \infty$, then U can be defined so as to be bounded and linear on H_X to H_Y; U is unique on G and U is an arbitrary bounded linear operator on $G^\perp$. Let $\mathfrak{Z}'$ be the span of $\mathfrak{Z}$ and U. Then U is efficient relative to g, h, $\mathfrak{Z}'$.

If $B = \infty$, one may consider whether the exclusion from Ω_1 of a null set may not make B finite. If it is impossible to make B finite by excluding a null set from Ω_1, one may consider whether a suitable change of topology may not make B finite. It does not seem appropriate to modify H_Y since h_ω is an element of H_Y for almost all ω and our object is to approximate h_ω in the topology of H_Y. We therefore consider the possibility of modifying H_X.

We say that a Hilbert space $H_{X'}$ is an *adequate substitute* (for H_X) if $g \in H_{X'\Omega} = H_{X'} \not\subset H_\Omega$, if each T in $\mathfrak{Z}$ is bounded and linear on $H_{X'}$ to H_Y, and if $B' < \infty$, where B' is defined as was B except that X is replaced by X'. If an adequate substitute $H_{X'}$ exists, it will be possible to enlarge $\mathfrak{Z}$ so that an efficient operator will exist (relative to $H_{X'}$ instead of H_X). The next section contains the construction of an adequate substitute in a particular case.

6. Illustration. Let H_Ω, H_X, H_Y be the spaces of functions absolute square integrable on $0 \leqq \omega < 1$, $0 \leqq x < 1$, $0 \leqq y < 1$ relative to Lebesgue measure, respectively. Then $\exp(2\pi i n\omega)$, $\exp(2\pi i n x)$, $\exp(2\pi i n y)$; $n = 0, \pm 1, \pm 2, \cdots$; are complete orthonormal sets in the respective spaces. Coordinates and matrices will refer to these orthonormal sets.

Let

$$g = g(x, \omega) = \sum_{n \neq 0} n^{-3} \exp [2\pi in (x + \omega)],$$

$$h = h(y, \omega) = \sum_{n \neq 0} n^{-2} \exp [2\pi in (y + \omega)].$$

Let $\mathfrak{Z}$ be the set of all bounded linear operators T on H_X to H_Y with diagonal matrices. Thus $\mathfrak{Z}$ consists of all operators of the form

$$T = \{\delta_{n,n'} s_n\}, \qquad \sup_n | s_n | < \infty, \qquad n, n' = 0, \pm 1, \pm 2, \cdots;$$

and

$$Tg = \sum_{n \neq 0} n^{-3} s_n \exp [2\pi in (y + \omega)].$$

Hence $L^0 = \mathfrak{Z}g$ surely includes all elements $\exp [2\pi in (y + \omega)]$, $n \neq 0$, and L consists of all elements in H_Y of the form

$$\sum_n t_n \exp [2\pi in (y + \omega)], \qquad \sum_n | t_n |^2 < \infty, \qquad t_0 = 0$$

Hence $h \in L$, $\theta = h$, and an operator U is efficient only if $Ug = h$. Suppose that U is such an operator.

It can be shown that G is the orthogonal complement of the line spanned by 1 in H_X. Also that U is unbounded on G^0 to H_Y, since its matrix on G to H_Y must be $\{\mu\delta_{n,\mu}\}$, $n = 0, \pm 1, \pm 2, \cdots$; $\mu = \pm 1, \pm 2, \cdots$. Hence no efficient operator exists.

In the present instance an adequate substitute for H_X can be obtained, for example, by using the topology of the image UG as that of $H_{X'}$, as follows. Let $H_{X'}$ be the space of elements u of G such that $\sum_n n^2 | u_n |^2 < \infty$ (and $u_0 = 0$), where $\{u_n\}$ are the coordinates of u in H_X, the new inner product being

$$(u, v)_{X'} = \sum_n n^2 u_n \bar{v}_n, \qquad u = \{u_n\} \in H_{X'}, \qquad v = \{v_n\} \in H_{X'};$$

and addition and scalar multiplication being as in H_X. Then $H_{X'}$ is a Hilbert space and an adequate substitute for H_X. Let $\mathfrak{Z}'$ be the set of all bounded linear operators on $H_{X'}$ to H_Y with diagonal matrices. Then U is efficient relative to $g, h, \mathfrak{Z}'$.

The space $H_{X'}$ gives more weight to higher frequencies than does H_X. This is related to the fact that, except for a constant factor, $h(x)$ is the derivative of $g(x)$ with respect to x.

7. A condition that L^0 be closed. In this section (and only here) we assume that g is *essentially dense* in the following sense: For any set Ω_2 which is almost all of Ω_1, the span in H_X of g_ω for all ω in Ω_2 is dense in G (Cf. §5).

Let $\mathfrak{B}$ be the Banach space of bounded linear operators on G to H_Y. Denote the norm of an operator V in $\mathfrak{B}$ by $\| V \|$.

Theorem 2. *Suppose that $\mathfrak{Z}$ is closed in $\mathfrak{B}$. Then L^0 is closed in $H_{Y\Omega}$ if and only if a positive constant k exists such that*

$$\| Tg \|_{Y\Omega} \geqq k \| T \| \tag{15}$$

for all T in $\mathfrak{Z}$.

If L^0 is closed, then $L = L^0$ and an efficient operator surely exists for all h in $H_{r\Omega}$, by Theorem 1. The hypothesis that $\mathfrak{Z}$ be closed in $\mathfrak{B}$ is not unduly restrictive. For if $\mathfrak{Z}$ were not closed, we could replace $\mathfrak{Z}$ by its closure in $\mathfrak{B}$ at the beginning of our analysis.

It seems simpler to give the proof of the above theorem than to adduce references. The proof is well known, except perhaps for the use of the essential denseness of g.

Proof. Necessity. Suppose that $L^0 = \mathfrak{Z}g$ is closed. Then

$$T \to Tg, \qquad T \in \mathfrak{Z}, \tag{16}$$

is a map of the Banach space $\mathfrak{Z}$ onto all of the Banach space L^0. The map is linear. The map is continuous, since

$$\| Tg \|_{r\Omega} \leqq \| T \| \, \| g \|_{r\Omega}$$

by Lemma 3. The map is one-to-one; for if $T_1 g = T_2 g$, $T_1, T_2 \in \mathfrak{Z}$, then $(T_2 - T_1)g = 0 \in H_{r\Omega}$ and $(T_2 - T_1)g_\omega = 0 \in H_Y$ for almost all ω, by Lemma 3. Hence $T_2 - T_1 = 0$ on G, since g is essentially dense.

By Banach's theorem [1, p. 41] the inverse map of (16) must be linear and continuous, hence bounded. Hence $B > 0$ exists such that

$$\| T \| \leqq B \, \| Tg \|_{r\Omega};$$

that is, (15) holds with $k = B^{-1}$.

Sufficiency. Suppose that $\{T^\nu g\}$, $\nu = 1, 2, \cdots$, $T^\nu \in \mathfrak{Z}$, is a Cauchy sequence in $H_{r\Omega}$. Then $\| (T^\nu - T^{\nu'})g \|_{r\Omega} \to 0$ as $\nu, \nu' \to \infty$. Then $\| T^\nu - T^{\nu'} \| \to 0$, by (15). Since $\mathfrak{Z}$ is closed, T exists such that $\| T^\nu - T \| \to 0$, $T \in \mathfrak{Z}$. Hence $\| T^\nu g - Tg \|_{r\Omega} \to 0$, by Lemma 3. Hence $T^\nu g \to Tg \in L^0$. Hence L^0 is closed.

8. Strongly efficient operators. These were defined in §3. The principle result of the present and the next section is Theorem 4 below.

Since

$$E \| e \|_Y^2 = \int_Y E \mid e(y) \mid^2 dm(y),$$

(2) implies (1); hence *a strongly efficient operator is surely efficient.* Therefore it is necessary for the existence of a strongly efficient operator that an efficient operator exist. If T_1 is an efficient operator, then $T_1 g$ and hence $e_1 = T_1 g - h$ are unique, by Theorem 1; hence either T_1 is strongly efficient or else no strongly efficient operator exists.

Likewise since (4) and (5) imply (3), *an operator within ϵ of strong efficiency ($\epsilon > 0$) is surely within ϵ of efficiency.*

Assume that L^0 contains an element other than 0. (Otherwise $L^0 = \mathfrak{Z}g = 0$, and 0 would be strongly efficient for lack of competition.)

Let ψ^λ, $\lambda \in \Lambda$, be an orthonormal set of elements of L^0 which are a basis for L [6; pp. 18, 19]. For each λ in Λ and each y with null exceptions, $\psi_y^\lambda \in H_\Omega$. Since Λ is countable, there is a set Y_1 which is almost all of Y such that $\psi_y^\lambda \in H_\Omega$

for each y in Y_1 and each λ in Λ. *For each y in Y_1, let M_y^0 be the span in H_Ω of ψ_y^λ, $\lambda \in \Lambda$; and let M_y be the closure in H_Ω of M_y^0.*

Lemma 5. *If $u \in L$, then $u_y \in M_y$ for almost all y.*

Proof. Since u is a limit point of L^0, a sequence $\{u^\nu\}$, $\nu = 1, 2, \cdots$, exists such that $\{u^\nu\}$ converges in $H_{T\Omega}$ to u and u^ν, for each ν, is a finite linear combination of the elements ψ^λ, $\lambda \in \Lambda$. For each ν and each y in Y_1, then, $u_y^\nu \in M_y^0$. By Lemma 2 a subsequence $\{u^{\nu_n}\}$ of $\{u^\nu\}$ exists such that $\{u_y^{\nu_n}\}$ converges in H_Ω to u_y for almost all y. Hence u_y, for almost all y, is in the closure of M_y^0, that is, in M_y. This completes the proof.

Put

$$\sigma_y = \mathrm{Proj}_{M_y} h_y ;$$

thus σ_y is the projection in H_Ω of h_y on M_y and is defined for almost all y.

Theorem 3. *An operator T_1 in $\mathfrak{Z}$ is strongly efficient if*

$$(T_1 g)_y = \sigma_y \qquad \text{for almost all } y. \tag{17}$$

An operator T_1 in $\mathfrak{Z}$ is within ϵ of strong efficiency ($\epsilon > 0$) if a function $\zeta = \zeta_y$ exists such that

$$E \mid (T_1 g)_y - \sigma_y \mid^2 \leqq \mid \zeta_y \mid^2 \qquad \text{for almost all } y \tag{18}$$

and

$$\zeta \in H_T, \qquad \| \zeta \|_T < \epsilon.$$

Remark. Our next theorem will imply that the condition (17) is necessary for strong efficiency as well as sufficient. Also that, if operators within ϵ of strong efficiency exist for all positive ϵ, the above condition for being within ϵ of strong efficiency is necessary as well as sufficient.

Proof of Theorem 3. Put

$$\rho_y = \sigma_y - h_y ; \tag{19}$$

then ρ_y is defined for almost all y. Also

$$\rho_y \perp M_y \qquad \text{for almost all } y, \tag{20}$$

since ρ_y is of the form $u - \mathrm{Proj}_{M_y} u$.

For any operator T in $\mathfrak{Z}$, put

$$e = Tg - h.$$

For almost all y,

$$\begin{aligned} e_y &= \rho_y + (e_y - \rho_y); \\ e_y - \rho_y &= (Tg)_y - \sigma_y \in M_y ; \end{aligned} \tag{21}$$

by Lemma 5 since $Tg \in L$. Thus ρ_y and $e_y - \rho_y$ are in orthogonal components of H_Ω and

$$E \mid e_y \mid^2 = E \mid \rho_y \mid^2 + E \mid e_y - \rho_y \mid^2, \tag{22}$$

$$E \mid e_y \mid^2 \geqq E \mid \rho_y \mid^2 \tag{23}$$

for almost all y.

Suppose that (17) holds. Put

$$e_1 = T_1 g - h.$$

Then $e_1(y) = \rho_y$ for almost all y, by (19). Hence (23) implies (2) and T_1 is strongly efficient.

Suppose that (18) holds. Then (22) and (21) with T replaced by T_1 imply that

$$E \mid e_1(y) \mid^2 = E \mid \rho_y \mid^2 + E \mid (T_1 g)_y - \sigma_y \mid^2 \qquad \text{for almost all } y.$$

Hence, by (23),

$$E \mid e(y) \mid^2 + \mid \zeta_y \mid^2 \geqq E \mid \rho_y \mid^2 + \mid \zeta_y \mid^2$$

$$= E \mid e_1(y) \mid^2 - E \mid (T_1 g)_y - \sigma_y \mid^2 + \mid \zeta_y \mid^2 \geqq E \mid e_1(y) \mid^2 \quad \text{for almost all } y.$$

Thus (5) is established and T_1 is within ϵ of strong efficiency, as was to be shown.

9. Efficiency sometimes implies strong efficiency. Lemma 6. *A necessary and sufficient condition that*

$$\theta_y = \sigma_y \qquad \textit{for almost all } y, \tag{24}$$

where

$$\theta = \mathrm{Proj}_L h, \qquad \sigma_y = \mathrm{Proj}_{M_y} h_y,$$

is that

$$\eta_y = \theta_y - h_y \perp M_y \qquad \textit{for almost all } y. \tag{25}$$

Proof. Since $\theta \in L$, $\theta_y - \sigma_y \in M_y$ for almost all y, by Lemma 5. Hence (24) is equivalent to the condition

$$\theta_y - \sigma_y \perp M_y \qquad \text{for almost all } y. \tag{26}$$

But

$$\eta_y = \rho_y + \theta_y - \sigma_y$$

by (19); and $\rho_y \perp M_y$ by (20). Hence (26) is equivalent to (25), and the lemma is proved.

Theorem 4. *If*

$$\theta_y = \sigma_y \qquad \textit{for almost all } y, \tag{27}$$

any operator within ϵ of efficiency ($\epsilon > 0$) is within ϵ of strong efficiency and any efficient operator is strongly efficient.

Conversely suppose that an operator within ϵ of strong efficiency exists for each positive ϵ. Then (27) holds. In particular, (27) holds if a strongly efficient operator exists.

Because of this theorem we shall say that *efficiency is strong* if and only if (27) holds.

Proof. The first part of the theorem is an immediate consequence of Theorem 3. Thus suppose that (27) holds. If T_1 in $\mathfrak{Z}$ is within ϵ of efficiency ($\epsilon > 0$),

$$E \| T_1 g - \theta \|_Y^2 < \epsilon^2$$

by Theorem 1. Put

$$\zeta^2 = E | T_1 g - \theta |^2, \qquad \zeta \geqq 0;$$

then $\zeta = \zeta(y)$ is measurable m and an element of H_Y. Also $\| \zeta \|_Y < \epsilon$, and condition (18) holds (with equality), by construction and (27). Hence T_1 is within ϵ of strong efficiency. Likewise if T_1 is efficient, condition (17) holds and T_1 is strongly efficient.

Conversely, assume that for each positive ϵ an operator within ϵ of strong efficiency exists. Assume that (27) does not hold. We shall derive a contradiction.

Note first that an element T_0 in $\mathfrak{Z}$ exists such that

$$q_0 = q_0(y) = E\eta(y)\bar{\gamma}_0(y) \neq 0, \tag{28}$$

where

$$\eta = \theta - h, \qquad \gamma_0 = T_0 g. \tag{29}$$

For if not, η_y would be normal in H_Ω to $(Tg)_y$ for almost all y, for all T in $\mathfrak{Z}$, since $q_0(y)$ is the inner product in H_Ω of $\eta(y)$ into $\gamma_0(y)$. Now M_y^0 is a span of elements $(Tg)_y$ for almost all y (Cf. §8). Hence η_y would be normal to M_y^0 and so to M_y for almost all y. Hence (24) = (27) would hold, by Lemma 6, contrary to our assumption that (27)' does not hold.

Choose a complex number c such that

$$\text{Real } \bar{c}q_0(y) > 0, \qquad y \in P, \tag{30}$$

where P is a measurable subset of Y of positive measure. This is possible. For if not, Real $\bar{c}q_0 \leqq 0$ almost everywhere for all complex numbers c. Hence Real $\bar{c}q_0 = 0$ almost everywhere for all c. (Otherwise Real $\bar{c}q_0 < 0$ and Real $-\bar{c}q_0 > 0$ somewhere for some c.) Hence $q_0 = 0$ almost everywhere, a contradiction of (28).

Put

$$\varphi = \theta - c\gamma_0,$$
$$\xi = \varphi - h = \eta - c\gamma_0.$$

Then

$$\begin{aligned} E | \xi |^2 &= E | \eta |^2 - 2 \text{ Real } E\eta\bar{c}\bar{\gamma}_0 + E | c\gamma_0 |^2 \\ &= E | \eta |^2 - 2 \text{ Real } \bar{c}q_0 + | c |^2 E | \gamma_0 |^2, \end{aligned}$$

$$\int_P E | \xi |^2 = \int_P E | \eta |^2 - 2 \text{ Real } \bar{c} \int_P q_0 + | c |^2 \int_P E | \gamma_0 |^2, \tag{31}$$

all integrations being relative to $dm(y)$. By (30),

$$\text{Real } \bar{c} \int_P q_0 > 0.$$

Now replace c by αc, where α is a positive number so small that the term in $\bar{c}$ in (31) dominates the term in $|c|^2$. Then (31) implies that

$$\int_P E|\xi|^2 < \int_P E|\eta|^2. \tag{32}$$

Let Q in Y be the set on which

$$E|\xi|^2 < E|\eta|^2.$$

Then Q is measurable; and (32) implies that $m(Q) > 0$. Put

$$\Delta^2 = \int_Q (E|\eta|^2 - E|\xi|^2) > 0, \tag{33}$$

$$B^2 = \int_Q E|\eta|^2 > 0, \qquad B > 0. \tag{34}$$

Choose ϵ so that

$$\epsilon > 0, \qquad \epsilon^2 < \Delta^2/8, \qquad \epsilon < \Delta^2/8B, \tag{35}$$

as is possible. Take T_1 within ϵ of strong efficiency, as is possible by hypothesis. Then T_1 is within ϵ of efficiency and $T_1 g$ is within ϵ of θ in $H_{Y\Omega}$, by Theorem 1. Take T_2 in $\mathfrak{I}$ so that $T_2 g$ is within ϵ of φ, as is possible since $\varphi \in L$ and $L^0 = \mathfrak{I}g$. Put

$$e_1 = T_1 g - h, \qquad e_2 = T_2 g - h.$$

Then

$$e_1 - \eta = T_1 g - \theta, \qquad e_2 - \xi = T_2 g - \varphi,$$

$$\int_Q E|e_1 - \eta|^2 = \int_Q E|T_1 g - \theta|^2 < \epsilon^2, \tag{36}$$

$$\int_Q E|e_2 - \xi|^2 = \int_Q E|T_2 g - \varphi|^2 < \epsilon^2, \tag{37}$$

since these inequalities would be true even with Q replaced by Y.

For the next paragraphs, let

$$\|v\|^2 = \int_Q E|v|^2. \tag{38}$$

We shall use the fact that

$$|\,\|v\|^2 - \|w\|^2\,| < \epsilon^2 + 2\epsilon \min(\|v\|, \|w\|) \tag{39}$$

whenever $\|v - w\| < \epsilon$ and $v, w \in H_{Y\Omega}$.

Since T_1 is within ϵ of strong efficiency,

$$E\,|\,e_1\,|^2 \leqq E\,|\,e_2\,|^2 + |\,\zeta\,|^2 \qquad \text{for almost all } y,$$

where

$$\zeta \,\varepsilon\, H_Y, \qquad \|\,\zeta\,\|_Y < \epsilon.$$

Hence for almost all y,

$$E\,|\,\eta\,|^2 - E\,|\,\xi\,|^2 \leqq |\,\zeta\,|^2 + (E\,|\,e_2\,|^2 - E\,|\,\xi\,|^2) - (E\,|\,e_1\,|^2 - E\,|\,\eta\,|^2).$$

Integrate over Q, and use (33), (38), (39), (37), (36), (34), (35) and the fact that $B = \|\,\eta\,\| > \|\,\xi\,\|$. Then

$$\Delta^2 < \epsilon^2 + |\,\|\,e_2\,\|^2 - \|\,\xi\,\|^2\,| + |\,\|\,e_1\,\|^2 - \|\,\eta\,\|^2\,| < \epsilon^2 + \epsilon^2 + 2\epsilon\,\|\,\xi\,\|$$
$$+ \epsilon^2 + 2\epsilon\,\|\,\eta\,\| < 3\epsilon^2 + 2\epsilon B + 2\epsilon B < 3\Delta^2/8 + 4\Delta^2/8 = 7\Delta^2/8.$$

This contradicts the fact that $\Delta^2 > 0$ and completes the proof.

Remark. *If $h \,\varepsilon\, L^0$, an efficient operator exists; also any efficient operator is exact and strongly efficient. If $h \,\epsilon\, L$, efficiency is strong.*

Proof. If $h \,\varepsilon\, L^0$, then $\theta = h = Tg$, $T \,\varepsilon\, \mathfrak{Z}$. So $e = Tg - h = 0$ and T is strongly efficient, by definition.

If $h \,\varepsilon\, L$, then $\theta = h$ and $h_y \,\varepsilon\, M_y$ for almost all y, by Lemma 5. Hence $\sigma_y = \mathrm{Proj}_{M_y}\, h_y = h_y = \theta_y$; and (27) holds.

Lemma 7. *Efficiency is strong if and only if*

$$h'_y \perp M_y \qquad \textit{for almost all } y,$$

where

$$h' \doteq \mathrm{Proj}_{L^\perp}\, h = h - \mathrm{Proj}_L\, h.$$

Proof. This lemma is a rephrasing of Lemma 6, since $h' = -\eta$.

10. A sufficient condition that efficiency be strong. A bounded linear operator T on H_X to H_Y is said to be *of finite rank* if the image TH_X of H_X under T is finite dimensional.

Theorem 5. *Suppose that $\mathfrak{Z}$ includes all operators of finite rank. Then efficiency is strong.*

The hypothesis of this theorem bears neither on g nor on h but solely on $\mathfrak{Z}$.

Proof. By Lemma 7, it is sufficient to show that the condition $w \perp L$ implies that

$$w_y \perp M_y \qquad \text{for almost all } y.$$

Let $u_j(y)$, $j \,\varepsilon\, J$, be a complete orthonormal set of elements of H_Y. Let $\varphi_\nu(\omega)$, $\nu \,\varepsilon\, N$, be a complete orthonormal set of elements of H_Ω. Then

$$u_j(y)\varphi_\nu(\omega), \qquad j \,\varepsilon\, J,\ \nu \,\varepsilon\, N,$$

is complete orthonormal set of elements of $H_{Y\Omega}$. Also

$$(\varphi_\nu\,,\,\varphi_{\nu'})_\Omega = E\varphi_\nu(\omega)\bar{\varphi}_{\nu'}(\omega) = \delta_{\nu,\nu'}\,, \qquad \nu,\,\nu' \,\epsilon\, N. \tag{40}$$

Coordinates in H_Y, H_Ω, $H_{Y\Omega}$ will refer to these orthonormal sets. As for H_X, we do not need an explicit notation for its basis. If $v \varepsilon H_X$, we denote the coordinates of v relative to a fixed coordinate system by v_k, $k \varepsilon K$. In $H_{X\Omega}$ we use the coordinate system based on the systems in H_X and H_Ω.

A bounded linear operator T on H_X to H_Y can be described by a matrix $\{t_{j,k}\}$, $j \varepsilon J$, $k \varepsilon K$, in the usual way. In particular, if v_k, $k \varepsilon K$, are the coordinates of v in H_X, then $\sum_k t_{j,k}v_k$, $j \varepsilon J$, are the coordinates of Tv in H_Y; and if $v_{k,\nu}$, $k \varepsilon K$, $\nu \varepsilon N$, are the coordinates of v in $H_{X\Omega}$, then $\sum_k t_{j,k}v_{k,\nu}$, $j \varepsilon J$, $\nu \varepsilon N$, are the coordinates of Tv in $H_{Y\Omega}$.

In the following paragraphs indices j, j' range over J, k ranges over K, ν, ν' range over N.

Let T^{j_0,k_0}, where j_0, k_0 are fixed and $j_0 \varepsilon J$, $k_0 \varepsilon K$, be the bounded linear operator on H_X to H_Y whose matrix is

$$\{\delta_{j,j_0}\delta_{k,k_0}\}.$$

Since T^{j_0,k_0} is of finite rank, it is an element of $\mathfrak{Z}$. (The only fact about $\mathfrak{Z}$ that we will use in the present proof is that $\mathfrak{Z}$ contains all operators T^{j_0,k_0}, $j_0 \varepsilon J$, $k_0 \varepsilon K$.)

Let w be any element of $H_{Y\Omega}$. Denote the coordinates of w by $w_{j,\nu}$. Denote the coordinates of g by $g_{k,\nu}$. Assume that $w \perp L$.

Then $w \perp Tg$ for all T in $\mathfrak{Z}$; in particular

$$(w, T^{j_0,k_0}g)_{Y\Omega} = E(w, T^{j_0,k_0}g)_Y = 0.$$

Now $T^{j_0,k_0}g$ has coordinates

$$\sum_k \delta_{j,j_0}\delta_{k,k_0}g_{k,\nu} = \delta_{j,j_0}g_{k_0,\nu}.$$

Hence

$$\sum_{j,\nu} w_{j,\nu}\bar{\delta}_{j,j_0}\bar{g}_{k_0,\nu} = 0,$$

$$\sum_{\nu} w_{j_0,\nu}\bar{g}_{k_0,\nu} = 0, \qquad j_0 \varepsilon J, \quad k_0 \varepsilon K. \tag{41}$$

Let ψ^λ, $\lambda \varepsilon \Lambda$, be the orthonormal set of elements of L^0 that generate L and that were used in §8 to define M_y^0. Then $w_y \perp M_y$ for almost all y if and only if

$$Ew(y)\bar{\psi}^\lambda(y) = 0 \qquad \text{for almost all } y, \text{ for all } \lambda \text{ in } \Lambda. \tag{42}$$

We now establish (42). Consider any fixed λ in Λ. An operator T exists such that

$$\psi^\lambda = Tg, \qquad T \varepsilon \mathfrak{Z}, \qquad T = \{t_{j,k}\}.$$

In the sense of convergence in the mean square relative to mp,

$$w = \sum_{j,\nu} w_{j,\nu}u_j(y)\varphi_\nu(\omega),$$

$$\psi^\lambda = Tg = \sum_{j,\nu}\sum_k t_{j,k}g_{k,\nu}u_j(y)\varphi_\nu(\omega).$$

Hence, in the sense of convergence in the mean absolute first power relative to mp,

$$w\bar{\psi}^{\lambda} = \sum_{\substack{j,\nu,\\ j',k,\nu'}} w_{j,\nu}\, u_j(y)\varphi_\nu(\omega)\bar{l}_{j',k}\,\bar{g}_{k,\nu'}\,\bar{u}_{j'}(y)\bar{\varphi}_{\nu'}(\omega).$$

Hence, in the sense of convergence in the mean absolute first power relative to m,

$$Ew\bar{\psi}^{\lambda} = \sum_{\substack{j,\nu,\\ j',k}} w_{j,\nu}\, u_j(y)\bar{l}_{j',k}\,\bar{g}_{k,\nu}\,\bar{u}_{j'}(y),$$

by (40). Hence $Ew\bar{\psi}^{\lambda} = 0$ in the sense of the mean absolute first power relative to m, by (41). The inversions above of orders of operations are permissible. Hence $Ew(y)\bar{\psi}^{\lambda}(y) = 0$ for almost all y, and (42) is established. This completes the proof.

The following example shows that efficiency need not be strong. Let H_Y be the two-dimensional space spanned by the orthonormal pair of real functions $u_1 = u_1(y)$, $u_2 = u_2(y)$. Let H_X be the one-dimensional space spanned by the unit vector v. Let H_Ω be the two-dimensional space spanned by the real orthonormal pair 1 and $\varphi = \varphi(\omega)$. Let $\mathfrak{I}$ be the set of all linear operators with matrices (relative to the bases v and u_1, u_2) of the form $\begin{Bmatrix} r \\ -r \end{Bmatrix}$, where r ranges over the real numbers. Let $g = v\varphi$, $h = (u_1 + u_2)\varphi$. Then the respective coordinates of g, h, Tg, and $e = Tg - h$ are

$$\{0 \quad 1\}, \quad \begin{Bmatrix} 0 & 1 \\ 0 & 1 \end{Bmatrix}, \quad \begin{Bmatrix} 0 & r \\ 0 & -r \end{Bmatrix}, \quad \begin{Bmatrix} 0 & r & -1 \\ 0 & -r & -1 \end{Bmatrix},$$

where

$$T = \begin{Bmatrix} r \\ -r \end{Bmatrix}.$$

Hence

$$e = (r-1)u_1\varphi - (r+1)u_2\varphi,$$

$$E \,\|\, e \,\|\, _Y^2 = (r-1)^2 + (r+1)^2,$$

$$E \,|\, e(y) \,|\, ^2 = |\, (r-1)u_1 - (r+1)u_2 \,|\, ^2.$$

Hence T is efficient if and only if $r = 0$, in which case $E \,\|\, e \,\|\, _Y^2 = 2$. The efficient operator is however not strongly efficient, as can be shown in detail and as is indicated by the fact that for almost all y, $E \,|\, e(y) \,|\, ^2$ has a minimum of zero at $r = (u_1 + u_2)/(u_1 - u_2)$.

REFERENCES

1. S. Banach, *Théorie des opérations linéaires*, Warszawa, 1932.
2. J. L. Doob, *Stochastic processes*, New York, 1953.
3. P. R. Halmos, *Measure theory*, New York, 1950.

4. A. N. KOLMOGOROV, *Interpolation und Extrapolation von stationären zufälligen Folgen*, Izvestiya Akad. Nauk SSSR. Ser. Mat., vol. 5 (1941) pp. 3–14 (Russian, German summary).
5. A. N. KOLMOGOROV, *On the proof of the method of least squares*, Uspehi Matem. Nauk (N. S.), vol. 1 (11) no. 1 (1946) pp. 57–70 (Russian).
6. F. J. MURRAY, *Linear transformations in Hilbert space*, Princeton, 1941.
7. F. J. MURRAY AND J. v. NEUMANN, *On rings of operators*, Annals of Math., vol. 37 (1936) pp. 116–229.
8. O. NIKODYM, *Sur un theorème de M. S. Zaremba concernant les fonctions harmoniques*, Journal de Math., (9) vol. 12 (1933) pp. 95–108.
9. J. v. NEUMANN, *Functional operators*, 2 volumes, Princeton, 1950.
10. A. SARD, *Approximation and variance*, Transactions of the Amer. Math. Soc., vol. 73 (1952) pp. 428–446.
11. R. SCHATTEN, *A theory of cross-spaces*, Princeton, 1950.
12. H. WEYL, *The method of orthogonal projection in potential theory*, Duke Math. Journal, vol. 7 (1940) pp. 411–444.
13. N. WIENER, *Extrapolation, interpolation, and smoothing of stationary time series* (with two appendices by N. Levinson), Cambridge, Mass., 1949.

QUEENS COLLEGE, FLUSHING, N. Y.

(Received July 1, 1954)

THE SOLUTION OF THE DIFFUSION EQUATION BY A HIGH ORDER CORRECT DIFFERENCE EQUATION

By Jim Douglas, Jr.

1. Introduction. The numerical solution of the boundary-value problem for the diffusion equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}, 0 \leqq x \leqq 1, 0 \leqq t \leqq T, \tag{1.1}$$

has been treated by means of several difference equations. Explicit methods have been studied in detail [5, 6, 7], and it has been found that small time steps are required for stability and convergence. Thus, the calculations are rather tedious whenever such methods are used.

Several implicit procedures have been proposed that are stable for any size time step. The simplest of these methods is the backwards difference equation [3, 9]

$$\begin{cases} \Delta_x^2 w_{i,n+1} = \dfrac{w_{i,n+1} - w_{in}}{\Delta t}, i = 1, \cdots, N-1, n \geqq 0, \\ w_{0,n+1} = u_{0,n+1}, w_{N,n+1}, w_{i0} = u_{i0}, \end{cases} \tag{1.2}$$

where w_{in} is the value of the function at the lattice point $(i\Delta x, n\Delta t)$, $N\Delta x = 1$, and

$$\Delta_x^2 w_{in} = (w_{i-1,n} - 2w_{in} + w_{i+1,n})/(\Delta x)^2. \tag{1.3}$$

It has been shown [3] that w_{in} converges uniformly to u_{in} with their difference being $u_{in} - w_{in} = O(\Delta t)$ if the ratio $\Delta t/(\Delta x)^2$ is held fixed and the boundary conditions are sufficiently smooth. Inasmuch as it can be shown easily that

$$\Delta_x^2 u_{i,n+1} = \frac{u_{i,n+1} - u_{in}}{\Delta t} + g_{in},$$

where $g_{in} = O((\Delta x)^2 + \Delta t) = O(\Delta t)$, we see that the total truncation error and the elemental error g_{in} are of the same order.

Perhaps the next simplest implicit scheme is the difference equation proposed by Crank and Nicolson [2, 8]:

$$\Delta_x^2 w_{i,n+1} + \Delta_x^2 w_{in} = 2\frac{w_{i,n+1} - w_{in}}{\Delta t}. \tag{1.5}$$

In this case the elemental truncation is $O((\Delta x)^2 + (\Delta t)^2) = O(\Delta t)$. Here the elemental error is basically unbalanced, as the error in the x-direction is of a larger order of magnitude than that in the t-direction. It has been shown [8] that the solution converges for any ratio $\Delta t/(\Delta x)^2$ with the total error being $O(\Delta t)$ as above; however, the constant will be smaller because of the smaller individual errors.

From the results for these two methods it can be conjectured that the total truncation error will be of the same order of magnitude as that of the elemental truncation error. A refinement of the Crank-Nicolson method is proposed in this report for which the elemental error is $O((\Delta t)^2)$. It is shown that under somewhat severe restrictions of the initial and boundary conditions the total truncation error is also $O((\Delta t)^2)$.

Since this paper was submitted, S. H. Crandall [10] derived by another method a special case of our difference equation (2.8). For the particular choice $\lambda = 2(5)^{-\frac{1}{2}}$, he showed that the elemental truncation error was $O((\Delta t)^2)$; however, he did not treat the total truncation error.

2. Difference Equation. In order that the linear equations arising from the difference equation be easily solvable by a method equivalent to elimination [1], it is desirable that only these values of the unknown function corresponding to x_{i-1}, x_i, and x_{i+1} appear in the difference equation as written about x_i; this will be a guiding criterion. Also, it is convenient to limit the t-values to t_n and t_{n+1} to prevent excessive storage requirements in machine computation.

First, let

$$\frac{4\Delta t}{(\Delta x)^2} = \lambda \tag{2.1}$$

be constant throughout the discussion. Now,

$$\frac{\partial u_{i,n+\frac{1}{2}}}{\partial t} = \frac{u_{i,n+1} - u_{in}}{\Delta t} - \frac{1}{24}\frac{\partial^3 u^*}{\partial t^3}(\Delta t)^2, \tag{2.2}$$

where u^*_{iii} is evaluated somewhere between (x_i, t_n) and (x_i, t_{n+1}). The asterisk will be used to denote functions evaluated at points called for by Taylor's theorem. Also,

$$\begin{aligned}\frac{\partial^2 u_{i,n+\frac{1}{2}}}{\partial x^2} &= \frac{1}{2}\left[\frac{\partial^2 u_{i,n+1}}{\partial x^2} + \frac{\partial^2 u_{in}}{\partial x^2}\right] - \frac{1}{8}\frac{\partial^4 u^*}{\partial x^2 \partial t^2}(\Delta t)^2 \\ &= \tfrac{1}{2}[\Delta_x^2 u_{i,n+1} + \Delta_x^2 u_{in}] - \frac{1}{12}\frac{\partial^4 u_{i,n+\frac{1}{2}}}{\partial x^4}(\Delta x)^2 \\ &\quad - \frac{1}{96}\frac{\partial^5 u}{\partial t^2 \partial x^2}(\Delta x)^2(\Delta t)^2 - \frac{1}{360}\frac{\partial^6 u^*}{\partial x^6}(\Delta x)^4 - \frac{1}{8}\frac{\partial^4 u^*}{\partial x^2 \partial t^2}(\Delta t)^2.\end{aligned} \tag{2.3}$$

The term u_{xxxx} must be approximated, as it is multiplied by $(\Delta x)^2 = O(\Delta t)$. To use the fourth difference in x would violate the criterion taken earlier; however,

$$\frac{\partial^4 u}{\partial x^4} = \frac{\partial^3 u}{\partial t \partial x^2}, \tag{2.4}$$

and this may be approximated as below:

$$\begin{aligned}\frac{\partial^3 u}{\partial t \partial x^2} &= \frac{\Delta_x^2 u_{i,n+1} - \Delta_x^2 u_{in}}{\Delta t} - \frac{1}{12}\frac{\partial^5 u^*}{\partial t \partial x^4}(\Delta x)^2 \\ &\quad - \frac{1}{24}\frac{\partial^5 u^*}{\partial t^3 \partial x^2}(\Delta t)^2 - \frac{1}{180}\frac{\partial^6 u^*}{\partial x^6}\frac{(\Delta x)^4}{\Delta t}.\end{aligned} \tag{2.5}$$

Substituting (2.2), (2.3), and (2.5) into (1.1), and simplifying

$$(2.6)\qquad (1 - 2/3\lambda)\Delta_x^2 u_{i,n+1} + (1 + 2/3\lambda)\Delta_x^2 u_{in} = 2\frac{u_{i,n+1} - u_{in}}{\Delta t} + g_{in}$$

where

$$(2.7)\qquad g_{in} = O((\Delta t)^2)$$

if $u(x, t)$ is at least six times boundedly differentiable in the region $0 \leqq x \leqq 1$, $0 \leqq t \leqq T$.

Thus, a completely second order correct difference equation satisfied by the solution of the differential equation has been derived; a choice of our difference equation for the numerical solution of (1.1) is apparent:

$$(2.8)\qquad \begin{cases} w_{i0} = u_{i0} \\ (1 - 2/3\lambda)\Delta_x^2 w_{i,n+1} + (1 + 2/3\lambda)\Delta_x^2 w_{in} = 2\dfrac{w_{i,n+1} - w_{in}}{\Delta t}, \qquad n \geqq 0 \\ w_{0,n+1} = u_{0,n+1}, \; w_{N,n+1} = u_{N,n+1}. \end{cases}$$

3. Error Equation and Superposition. Let

$$(3.1)\qquad v_{in} = u_{in} - w_{in}.$$

The difference equation for the truncation error v_{in} can be obtained by subtracting (2.8) from (2.6):

$$(3.2)\qquad \begin{cases} v_{i0} = 0 \\ (1 - 2/3\lambda)\Delta_x^2 v_{i,n+1} + (1 + 2/3\lambda)\Delta_x^2 v_{in} = 2\dfrac{v_{i,n+1} - v_{in}}{\Delta t} + g_{in} \\ v_{0,n+1} = v_{N,n+1} = 0. \end{cases}$$

The solution of (3.2) may be facilitated by the use of the following principle of superposition. Let

$$(3.3)\qquad v_{in} = \sum_{k=0}^{n-1} v_{in}^{(k)},$$

where

$$(3.4)\qquad \begin{cases} v_{im}^{(k)} = 0, \qquad m \leqq k \\ (1 - 2/3\lambda)\Delta_x^2 v_{i,k+1}^{(k)} - \dfrac{2}{\Delta t} v_{i,k+1}^{(k)} = g_{ik}, \qquad m = k + 1 \\ (1 - 2/3\lambda)\Delta_x^2 v_{i,m+1}^{(k)} + (1 + 2/3\lambda)\Delta_x^2 v_{im}^{(k)} = 2\dfrac{v_{i,m+1}^{(k)} - v_{im}^{(k)}}{\Delta t}, \quad m \geqq k + 1. \end{cases}$$

Direct substitution shows that v_{in} as defined by (3.3) and (3.4) satisfies (3.2). Note that each of the functions $v_{i,n-k-1}^{(k)}$ satisfies the homogeneous difference equation related to (3.2) with a non-vanishing initial condition.

4. Solution of Homogeneous Equation. Let

$$\begin{cases} (1 - 2/3\lambda)\Delta_x^2 z_{i,n+1} + (1 + 2/3\lambda)\Delta_x^2 z_{in} = 2\dfrac{z_{i,n+1} - z_{in}}{\Delta t}, & n \geqq 0 \\ z_{0,n} = z_{N,n} = 0. \end{cases} \tag{4.1}$$

Consider as initial condition

$$z_{i0} = \sin \pi p x_i , \tag{4.2}$$

and assume a solution of the form

$$z_{in} = a_n \sin \pi p x_i . \tag{4.3}$$

Substituting into (4.1),

$$\frac{a_{n+1}}{a_n} = \frac{2 - \lambda(1 + 2/3\lambda)\sin^2 \pi p/2N}{2 + \lambda(1 - 2/3\lambda)\sin^2 \pi p/2N} = \rho_p . \tag{4.4}$$

The ratio ρ_p is less than one for any p, $0 < p < N$ and any $\lambda > 0$; thus, the integration method is stable for any size time step. The solution of (4.1) with the initial condition (4.2) then is

$$z_{in} = \rho_p^n \sin \pi p x_i . \tag{4.5}$$

As (4.1) is linear, its solution under the most general initial condition

$$z_{i0} = \sum_{p=1}^{N-1} c_p \sin \pi p x_i \tag{4.6}$$

is

$$z_{in} = \sum_{p=1}^{N-1} c_p \rho_p^n \sin \pi p x_i , \tag{4.7}$$

5. Solution of Error Equation. Expand each of functions $v_{i,k+1}^{(k)}$ in a Fourier series:

$$v_{i,k+1}^{(k)} = \sum_{p=1}^{N-1} c_p^{(k)} \sin \pi p x_i . \tag{5.1}$$

Then, by (4.7),

$$v_{in}^{(k)} = \sum_{p=1}^{N-1} c_p^{(k)} \rho_p^{n-k-1} \sin \pi p x_i . \tag{5.2}$$

By (3.3),

$$v_{in} = \sum_{k=0}^{n-1} \sum_{p=1}^{N-1} c_p^{(k)} \rho_p^{n-k-1} \sin \pi p x_i . \tag{5.3}$$

The asymptotic behavior of the error as the lattice is refined depends on that of the Fourier coefficients $c_p^{(k)}$ and the stability ratios ρ_p. Lemmas giving bounds for each will be developed before treating (5.3).

6. Fourier Coefficients. It is easy to see that if

$$g_{ik} = \sum_{p=1}^{N-1} a_p^{(k)} \sin \pi p x_i, \tag{6.1}$$

then

$$v_{i,k+1}^{(k)} = -\sum_{p=1}^{N-1} \frac{\Delta t a_p^{(k)}}{2 + (\lambda - 2/3) \sin^2 \pi p/2N} \sin \pi p x_i. \tag{6.2}$$

Now,

$$a_p^{(k)} = \frac{2}{N} \sum_{i=1}^{N-1} g_{ik} \sin \pi p x_i \tag{6.3}$$

is closely related to the Fourier coefficient

$$A_p^{(k)} = 2 \int_0^1 g_k(x) \sin \pi p x \, dx \tag{6.4}$$

where $g_k(x)$ is obtained by translating the lattice in the x-direction. It is easy to see from (2.6) that $g_k(x)$ possesses a bounded derivative with respect to x. The following lemma then applies:

Lemma 1. *If $u(x, t)$ is at least six times boundedly differentiable in $0 \leqq x \leqq 1$, $0 \leqq t \leqq T$, then there exists a constant A depending only on u and these derivatives such that*

$$| A_p^{(k)} - a_p^{(k)} | < A(\Delta t)^2 p/N. \tag{6.5}$$

Proof. The two dimensional analogue of this lemma has been proved previously [4]; this proof is an obvious simplification of the previous one and will be omitted. The factor $(\Delta t)^2$ comes from (2.7).

The conditions on $g_k(x)$ also insure that there exists B depending only on u and its derivatives of no higher than seventh order such that

$$| A_p^{(k)} | < B(\Delta t)^2/p. \tag{6.6}$$

If $C = \max(A, B)$,

$$| a_p^{(k)} | < C(1/p + p/N)(\Delta t)^2. \tag{6.7}$$

By (6.2),

$$| c_p^{(k)} | < C(1/p + p/N)(\Delta t)^2. \tag{6.8}$$

7. Stability Ratios. Consider the function

$$\begin{aligned} \rho = \rho(\xi) &= \frac{2 - (\lambda + 2/3)\xi}{2 + (\lambda - 2/3)\xi}, \qquad 0 \leqq \xi \leqq 1, \\ &= 1 - \frac{\lambda \xi}{1 + 1/2(\lambda - 2/3)\xi} \end{aligned} \tag{7.1}$$

As $\rho'(\xi) > 0$ for $0 \leqq \xi \leqq 1$, there is at most one root, say ξ_0, of $\rho(\xi) = 0$ in

$0 \leqq \xi \leqq 1$. For $0 \leqq \xi \leqq \xi_0$,

$$|\rho(\xi)| \leqq 1 - \frac{\lambda}{\eta}\xi, \tag{7.2}$$

where $\eta = \max(1, 1/6(3\lambda + 2))$. From ξ_0 to 1, $|\rho(\xi)|$ increases from 0 to $|(4 - 3\lambda)/(4 + 3\lambda)| < 1$. Thus,

$$|\rho(\xi)| \leqq 1 - (1 - |\rho(1)|)\xi \tag{7.3}$$

for $\xi_0 \leqq \xi \leqq 1$. Let

$$\delta = \min(\lambda/\eta, 1 - |\rho(1)|). \tag{7.4}$$

Then

$$|\rho(\xi)| \leqq 1 - \delta\xi, 0 \leqq \xi \leqq 1. \tag{7.5}$$

Lemma 2. *There exists a constant $D > 0$ such that*

$$\sum_{k=0}^{\infty} |\rho_p|^k \leqq DN^2/p^3. \tag{7.6}$$

Proof: From (7.5),

$$\sum_{k=0}^{\infty} |\rho(\xi)|^k < \frac{1}{\delta\xi}.$$

Now, if $\xi = \sin^2 \pi p/2N$, $\xi \geqq \pi^2 p^2/16N^2$. Thus, D may be taken equal to $16/\pi^2\delta$. q.e.d.

8. Convergence Theorem. The results of the last two sections can be applied to (5.3) as below.

$$\begin{aligned} |v_{in}| &\leqq \sum_{k=0}^{n-1} \sum_{p=1}^{N-1} |c_p^{(k)}| |\rho_p|^{n-k-1} \\ &< C(\Delta t)^3 \sum_{p=1}^{N-1} (1/p + p/N) \sum_{k=0}^{\infty} |\rho_p|^k \\ &< CDN^2(\Delta t)^3 \sum_{p=1}^{N-1} (1/p^3 + 1/Np). \end{aligned} \tag{8.1}$$

As

$$\begin{cases} \sum_{p=1}^{\infty} 1/p^3 < 3 \\ \sum_{p=1}^{N-1} 1/Np \approx \log N/N < 1, \end{cases} \tag{8.2}$$

there exists $E = 4CD$ such that

$$|v_{in}| < EN^2(\Delta t)^3 = \lambda E(\Delta t)^2/4. \tag{8.3}$$

This result may be collected in the following convergence theorem.

Theorem. *If there exists a solution $u(x, t)$ of the boundary value problem associated with the differential equation* (1.1) *and if $u(x, t)$ is at least seven times boundedly differentiable in the region $0 \leqq x \leqq 1$, $0 \leqq t \leqq T$ then the solution of the difference equation* (2.9) *converges uniformly to $u(x, t)$ in the region and the truncation error is $O((\Delta t)^3)$.*

9. Practical Aspects. If the method of [1] is used to solve the linear equations occurring at each time step, it can be seen easily that $7(N-1)$ calculations per time step are required by the backwards difference formula, $9(N-1)$ by the Crank-Nicolson method, and $10(N-1)$ by the method presented here. Thus, very little additional calculation per time step would be required even if the same number of subdivisions were used for the new method; however, it is obvious from the convergence rates for the three methods that both fewer x-increments and fewer time steps would be required to complete the numerical solution given here. Consequently, the procedure developed above is considerably superior to the previously treated implicit methods.

BIBLIOGRAPHY

1. Bruce, G. H., D. W. Peaceman, H. H. Rachford, and J. D. Rice. Calculations of unsteady-state gas flow through porous media, Trans. AIME **198** (1953) pp. 79–91.
2. Crank, J. and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type, Proc. Camb. Phil. Soc. **43** (1947) pp. 50–67.
3. Douglas, J. On the numerical integration of quasi-linear parabolic differential equations, to appear in the Pacific Journal of Mathematics.
4. ———, On the numerical integration of $U_{xx} + U_{yy} = U_t$ by implicit methods, Jour. Soc. Ind. and App. Math. **3** (1955) pp. 42–65.
5. Hildebrand, F. B. On the convergence of numerical solutions of the heat-flow equation, Jour. of Math. and Phys. XXXI (1952) pp. 35–41.
6. Juncosa, M. L. and D. Young. On the convergence of a solution of a difference equation to a solution to the equation of diffusion, Proc. Amer. Math. Soc. **5** (1954) pp. 168–174.
7. ———, On the order of convergence of solutions of a difference equation to a solution of the diffusion equation, Jour. Soc. Ind. and App. Math. **1** (1953) pp. 111–135.
8. ———, On the Crank-Nicolson procedure for parabolic partial differential equations, Abstract 240, Bull. Amer. Math. Soc. **60** (1954) p. 152.
9. O'Brien, G., M. Hyman, and S. Kaplan. A study of the numerical solution of partial differential equations, Jour. of Math. and Phys. XXIX (1951) pp. 223–251.
10. Crandall, S. H. An optimum implicit recurrence formula for the heat conduction equation, Quarterly of Applied Mathematics, XIII (1955) pp. 318–320.

Humble Oil & Refining Company

(Received December 28, 1954)

COEFFICIENTS FOR COMPLEX OSCULATORY INTERPOLATION OVER A CARTESIAN GRID

By Herbert E. Salzer

When an analytic function $f(z)$ and its first derivative $f'(z)$ are known at any n points z_k in the complex plane, the general Hermite osculatory interpolation formula becomes [1], [2]

$$f(z) = \sum_k \{L_k^{(n)}(z)\}^2\{1 - 2L_k^{(n)\prime}(z_k)(z - z_k)\}f(z_k) + \sum_k \{L_k^{(n)}(z)\}^2(z - z_k)f'(z_k) + R_{2n}(z), \tag{1}$$

where

$$L_k^{(n)}(z) = \prod_j'\{(z - z_j)/(z_k - z_j)\}, \qquad j = k \text{ absent from } \prod', \tag{2}$$

and

$$R_{2n}(z) = \frac{\prod_k \{(z - z_k)^2\}}{2\pi i}\int_C \frac{f(t)}{(t - z)\prod_k \{(t - z_k)^2\}}\,dt, \tag{3}$$

the $\prod_k$ denoting the product over all n points z_k. Here C is any contour on and within which $f(t)$ is analytic and which surrounds the $n + 1$ points z_k and z. The right member of (1) without the remainder term $R_{2n}(z)$ is the unique polynomial of the $(2n - 1)^{\text{th}}$ degree in z which coincides with $f(z)$ and $f'(z)$ at $z = z_k$. Some idea as to the magnitude of $R_{2n}(z)$ may be had from considering it as a limiting confluent form of the remainder in the ordinary Lagrangian formula over the $2n$ points z_l, the latter being expressible according to a theorem due to Paul Montel [3], as $\prod_l(z - z_l)f^{(2n)}(\alpha)/(2n)!$, where $f^{(2n)}(\alpha)$ is some point within the domain of convexity corresponding to the domain covered by $f^{(2n)}(t)$ when t covers a convex region containing the points z_l and z. In practice one could probably get a good idea as to the magnitude of $R_{2n}(z)$ by taking $f^{(2n)}(\alpha)$ for α equal to any z_l, or even to z, so that one would expect the approximate equality

$$R_{2n}(z) \sim \prod_k\{(z - z_k)^2\}f^{(2n)}(z_{k'})/(2n)!, \tag{3'}$$

the $z_{k'}$ denoting any z_k or z.

The present tables are intended to facilitate the use of (1) for complex osculatory interpolation for analytic functions $f(z)$ that are tabulated over a Cartesian grid either with their first derivatives $f'(z)$ or wherever the $f'(z)$ is easy to calculate. Even though tables for complex Cartesian interpolation based upon Lagrange's interpolation formula employing only the functional values $f(z_k)$ are available in published form [4], [5], these present tables should be used wherever it is convenient due to the much greater accuracy of the osculatory interpolation formula. Often the user will find the grid length h to be too large for the desired accuracy using the ordinary Lagrangian coefficients, but not too large when osculatory interpolation is employed. These tables should prove helpful in inter-

polation for many tables of higher mathematical functions for complex arguments in Cartesian form, viz., 1) Bessel functions of the first or second kind, $J_0(z)$ or $Y_0(z)$ tabulated with $J_1(z) = -J_0'(z)$ or $Y_1(z) = -Y_0'(z)$ respectively, or linear combinations of them, 2) $\log \Gamma(z)$ whenever its derivative $\psi(z)$ is available, 3) probability integral $\int_0^z e^{-u^2}\,du$ [6] with its integrand e^{-z^2}, 4) integrals of more elementary functions where the integrand is easy to calculate, like $\int_z^\infty (e^{-u}/u)\,du$ [7], and 5) solutions of important linear differential equations which are given with their first derivatives [8].

To interpolate for $f(z)$ when one has $f(z_k)$ and $f'(z_k)$ at arguments $z_k \equiv z_0 + kh$ arranged in a Cartesian grid of length h, one first chooses z_0 as the lower left-hand corner point of the square whose four corners are $z_0 \equiv z_0$, $z_1 \equiv z_0 + h$, $z_i \equiv z_0 + ih$, $z_{1+i} \equiv z_0 + (1 + i)h$ and which contains $z \equiv z_0 + Ph$, $P = p + iq$, either inside or upon a side of the square, so that $0 \leqq p \leqq 1$, $0 \leqq q \leqq 1$. Then (1) is expressible in the form

$$f(z_0 + Ph) = \sum_k \{A_k^{(n)}(P)f(z_k) + hB_k^{(n)}(P)f'(z_k)\} + R_{2n}, \tag{4}$$

where the coefficients $A_k^{(n)}(P)$ and $B_k^{(n)}(P)$ are polynomials of the $(2n - 1)^{\text{th}}$ degree in P. These present tables give the exact values of $A_k^{(n)}(P)$ and $B_k^{(n)}(P)$ for n corresponding to the two-, three-, and four-point cases based upon the choices of z_k as z_0, z_1, z_0, z_1, z_i, and z_0, z_1, z_i, z_{1+i} respectively, for $P = p + iq$ for both $p, q = 0(.1)1$. When written explicitly for each n, (4) is seen to become for $n = 2$ (cubic accuracy),

$$\begin{aligned} f(z_0 + Ph) \sim{}& A_0^{(2)}(P)f(z_0) + A_1^{(2)}(P)f(z_1) \\ &+ h\{B_0^{(2)}(P)f'(z_0) + B_1^{(2)}(P)f'(z_1)\}; \end{aligned} \tag{4'}$$

for $n = 3$ (quintic accuracy),

$$\begin{aligned} f(z_0 + Ph) \sim{}& A_0^{(3)}(P)f(z_0) + A_1^{(3)}(P)f(z_1) + A_i^{(3)}(P)f(z_i) \\ &+ h\{B_0^{(3)}(P)f'(z_0) + B_1^{(3)}(P)f'(z_1) + B_i^{(3)}(P)f'(z_i)\}; \end{aligned} \tag{4''}$$

and for $n = 4$ (septic accuracy),

$$\begin{aligned} f(z_0 + Ph) \sim{}& A_0^{(4)}(P)f(z_0) + A_1^{(4)}(P)f(z_1) + A_i^{(4)}(P)f(z_i) + A_{1+i}^{(4)}(P)f(z_{1+i}) \\ &+ h\{B_0^{(4)}(P)f'(z_0) + B_1^{(4)}(P)f'(z_1) + B_i^{(4)}(P)f'(z_i) + B_{1+i}^{(4)}(P)f'(z_{1+i})\}. \end{aligned} \tag{4'''}$$

In the Tables of Coefficients only $A_0^{(2)}(P)$, $B_0^{(2)}(P)$, $A_0^{(3)}(P)$, $A_1^{(3)}(P)$, $B_0^{(3)}(P)$, $B_1^{(3)}(P)$, $A_0^{(4)}(P)$ and $B_0^{(4)}(P)$ are tabulated explicitly since the other coefficients $A_k^{(n)}(P)$ and $B_k^{(n)}(P)$ are read from these tables in view of the following simple relations:

$$A_1^{(2)}(p + iq) = \overline{A_0^{(2)}(1 - p + iq)}, \tag{5}$$

$$B_1^{(2)}(p + iq) = -\overline{B_0^{(2)}(1 - p + iq)}; \tag{5'}$$

TABLES OF COEFFICIENTS $A_r^{(n)}(P)$ AND $B_r^{(n)}(P)$

Two-Point: $A_0^{(2)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	$4.000 - 2.000i$	$3.372 - 2.540i$	$2.696 - 2.960i$	$1.984 - 3.260i$	$1.248 - 3.440i$	$.500 - 3.500i$
0.9	$3.430 - 1.458i$	$2.916 - 1.944i$	$2.354 - 2.322i$	$1.756 - 2.592i$	$1.134 - 2.754i$	$.500 - 2.808i$
0.8	$2.920 - 1.024i$	$2.508 - 1.456i$	$2.048 - 1.792i$	$1.552 - 2.032i$	$1.032 - 2.176i$	$.500 - 2.224i$
0.7	$2.470 - .686i$	$2.148 - 1.064i$	$1.778 - 1.358i$	$1.372 - 1.568i$	$.942 - 1.694i$	$.500 - 1.736i$
0.6	$2.080 - .432i$	$1.836 - .756i$	$1.544 - 1.008i$	$1.216 - 1.188i$	$.864 - 1.296i$	$.500 - 1.332i$
0.5	$1.750 - .250i$	$1.572 - .520i$	$1.346 - .730i$	$1.084 - .880i$	$.798 - .970i$	$.500 - 1.000i$
0.4	$1.480 - .128i$	$1.356 - .344i$	$1.184 - .512i$	$.976 - .632i$	$.744 - .704i$	$.500 - .728i$
0.3	$1.270 - .054i$	$1.188 - .216i$	$1.058 - .342i$	$.892 - .432i$	$.702 - .486i$	$.500 - .504i$
0.2	$1.120 - .016i$	$1.068 - .124i$	$.968 - .208i$	$.832 - .268i$	$.672 - .304i$	$.500 - .316i$
0.1	$1.030 - .002i$	$.996 - .056i$	$.914 - .098i$	$.796 - .128i$	$.654 - .146i$	$.500 - .152i$
0.0	$1.000 - .000i$	$.972 - .000i$	$.896 - .000i$	$.784 - .000i$	$.648 - .000i$	$.500 - .000i$

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	$-.248 - 3.440i$	$-.984 - 3.260i$	$-1.696 - 2.960i$	$-2.372 - 2.540i$	$-3.000 - 2.000i$
0.9	$-.134 - 2.754i$	$-.756 - 2.592i$	$-1.354 - 2.322i$	$-1.916 - 1.944i$	$-2.430 - 1.458i$
0.8	$-.032 - 2.176i$	$-.552 - 2.032i$	$-1.048 - 1.792i$	$-1.508 - 1.456i$	$-1.920 - 1.024i$
0.7	$.058 - 1.694i$	$-.372 - 1.568i$	$-.778 - 1.358i$	$-1.148 - 1.064i$	$-1.470 - .686i$
0.6	$.136 - 1.296i$	$-.216 - 1.188i$	$-.544 - 1.008i$	$-.836 - .756i$	$-1.080 - .432i$
0.5	$.202 - .970i$	$-.084 - .880i$	$-.346 - .730i$	$-.572 - .520i$	$-.750 - .250i$
0.4	$.256 - .704i$	$.024 - .632i$	$-.184 - .512i$	$-.356 - .344i$	$-.480 - .128i$
0.3	$.298 - .486i$	$.108 - .432i$	$-.058 - .342i$	$-.188 - .216i$	$-.270 - .054i$
0.2	$.328 - .304i$	$.168 - .268i$	$.032 - .208i$	$-.068 - .124i$	$-.120 - .016i$
0.1	$.346 - .146i$	$.204 - .128i$	$.086 - .098i$	$.004 - .056i$	$-.030 - .002i$
0.0	$.352 - .000i$	$.216 - .000i$	$.104 - .000i$	$.028 - .000i$	$-.000 - .000i$

Two-Point: $B_0^{(2)}(P)$

q	p = 0.0	p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5
1.0	2.000 + .000i	1.781 − .370i	1.528 − .680i	1.247 − .930i	.944 − 1.120i	.625 − 1.250i
0.9	1.620 + .171i	1.458 − .162i	1.262 − .441i	1.038 − .666i	.792 − .837i	.530 − .954i
0.8	1.280 + .288i	1.169 − .008i	1.024 − .256i	.851 − .456i	.656 − .608i	.445 − .712i
0.7	.980 + .357i	.914 + .098i	.814 − .119i	.686 − .294i	.536 − .427i	.370 − .518i
0.6	.720 + .384i	.693 + .162i	.632 − .024i	.543 − .174i	.432 − .288i	.305 − .366i
0.5	.500 + .375i	.506 + .190i	.478 + .035i	.422 − .090i	.344 − .185i	.250 − .250i
0.4	.320 + .336i	.353 + .188i	.352 + .064i	.323 − .036i	.272 − .112i	.205 − .164i
0.3	.180 + .273i	.234 + .162i	.254 + .069i	.246 − .006i	.216 − .063i	.170 − .102i
0.2	.080 + .192i	.149 + .118i	.184 + .056i	.191 + .006i	.176 − .032i	.145 − .058i
0.1	.020 + .099i	.098 + .062i	.142 + .031i	.158 + .006i	.152 − .013i	.130 − .026i
0.0	.000 + .000i	.081 + .000i	.128 + .000i	.147 + .000i	.144 − .000i	.125 − .000i

q	p = 0.6	p = 0.7	p = 0.8	p = 0.9	p = 1.0
1.0	.296 − 1.320i	−.037 − 1.330i	−.368 − 1.280i	−.691 − 1.170i	−1.000 − 1.000i
0.9	.258 − 1.017i	−.018 − 1.026i	−.292 − .981i	−.558 − .882i	−.810 − .729i
0.8	.224 − .768i	−.001 − .776i	−.224 − .736i	−.439 − .648i	−.640 − .512i
0.7	.194 − .567i	.014 − .574i	−.164 − .539i	−.334 − .462i	−.490 − .343i
0.6	.168 − .408i	.027 − .414i	−.112 − .384i	−.243 − .318i	−.360 − .216i
0.5	.146 − .285i	.038 − .290i	−.068 − .265i	−.166 − .210i	−.250 − .125i
0.4	.128 − .192i	.047 − .196i	−.032 − .176i	−.103 − .132i	−.160 − .064i
0.3	.114 − .123i	.054 − .126i	−.004 − .111i	−.054 − .078i	−.090 − .027i
0.2	.104 − .072i	.059 − .074i	.016 − .064i	−.019 − .042i	−.040 − .008i
0.1	.098 − .033i	.062 − .034i	.028 − .029i	.002 − .018i	−.010 − .001i
0.0	.096 − .000i	.063 − .000i	.032 − .000i	.009 − .000i	−.000 − .000i

TABLES OF COEFFICIENTS $A_k^{(n)}(P)$ AND $B_k^{(n)}(P)$

Three-Point: $A_0^{(2)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	.00000 − .00000i	−.02632 + .06102i	−.05344 + .24064i	−.01116 + .51786i	.15872 + .85248i	.50000 + 1.18750i
0.9	.03772 − .04698i	.09720 + .05184i	.14968 + .20426i	.24616 + .39888i	.42804 + .60750i	.72472 + .78752i
0.8	.13984 − .14336i	.22652 − .01414i	.30720 + .12288i	.41608 + .26350i	.58016 + .38912i	.81684 + .46914i
0.7	.28656 − .23814i	.36104 − .09072i	.43532 + .03550i	.52920 + .14112i	.65768 + .21474i	.82856 + .23536i
0.6	.45568 − .29952i	.49536 − .14850i	.54304 − .03328i	.60652 + .04914i	.69120 + .09216i	.79768 + .08198i
0.5	.62500 − .31250i	.62168 − .17248i	.63456 − .07086i	.66184 − .00464i	.70172 + .02198i	.75000 + .00000i
0.4	.77472 − .27648i	.73220 − .15966i	.71168 − .07424i	.70416 − .01962i	.70304 + .00000i	.70172 − .02198i
0.3	.88984 − .20286i	.82152 − .11664i	.77620 − .04762i	.74008 − .00000i	.70416 + .01962i	.66184 + .00464i
0.2	.96256 − .11264i	.88904 − .05722i	.83232 − .00000i	.77620 + .04762i	.71168 + .07424i	.63456 + .07086i
0.1	.99468 − .03402i	.94136 − .00000i	.88904 + .05722i	.82152 + .11664i	.73220 + .15966i	.62168 + .17248i
0.0	1.00000 − .00000i	.99468 + .03402i	.96256 + .11264i	.88984 + .20286i	.77472 + .27648i	.62500 + .31250i

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	1.03968 + 1.45152i	1.78556 + 1.56114i	2.72384 + 1.42336i	3.81672 + .93798i	5.00000 + .00000i
0.9	1.15120 + .88434i	1.70568 + .83376i	2.36716 + .56438i	3.09304 + .00000i	3.81672 − .93798i
0.8	1.13152 + .46336i	1.51520 + .32438i	1.94208 + .00000i	2.36716 − .56438i	2.72384 − 1.42336i
0.7	1.04004 + .17478i	1.27832 + .00000i	1.51520 − .32438i	1.70568 − .83376i	1.78556 − 1.56114i
0.6	.91936 + .00000i	1.04004 − .17478i	1.13152 − .46336i	1.15120 − .88434i	1.03968 − 1.45152i
0.5	.79768 − .08198i	.82856 − .23536i	.81684 − .46914i	.72472 − .78752i	.50000 − 1.18750i
0.4	.69120 − .09216i	.65768 − .21474i	.58016 − .38912i	.42804 − .60750i	.15872 − .85248i
0.3	.60652 − .04914i	.52920 − .14112i	.41608 − .26350i	.24616 − .39888i	−.01116 − .51786i
0.2	.54304 + .03328i	.43532 − .03550i	.30720 − .12288i	.14968 − .20426i	−.05344 − .24064i
0.1	.49536 + .14850i	.36104 + .09072i	.22652 + .01414i	.09720 − .05184i	−.02632 − .06102i
0.0	.45568 + .29952i	.28656 + .23814i	.13984 + .14336i	.03772 + .04698i	−.00000 − .00000i

Three-Point: $A_1^{(q)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	.000000 + .000000i	−.015095 − .021165i	−.077440 − .066880i	−.204885 − .105795i	−.404480 − .101760i	−.671875 − .015625i
0.9	.006885 + .019845i	−.033560 + .022680i	−.097705 + .020815i	−.198000 + .036000i	−.337895 + .094185i	−.511240 + .223720i
0.8	.017920 + .061440i	−.026925 + .072225i	−.080320 + .092160i	−.149915 + .135195i	−.236160 + .218880i	−.331705 + .362565i
0.7	.024255 + .103635i	−.010640 + .113520i	−.045535 + .141005i	−.083880 + .194040i	−.123725 + .283575i	−.157120 + .421760i
0.6	.023040 + .132480i	.006445 + .138015i	−.006400 + .164800i	−.015345 + .216585i	−.016640 + .299520i	−.002335 + .420355i
0.5	.015625 + .140625i	.020280 + .141960i	.029435 + .163995i	.046240 + .208080i	.075645 + .277365i	.125000 + .375000i
0.4	.005760 + .126720i	.029815 + .125805i	.057920 + .141440i	.095625 + .174375i	.148480 + .226560i	.222635 + .299345i
0.3	−.002205 + .094815i	.035200 + .093600i	.076805 + .101785i	.129960 + .121320i	.200215 + .154755i	.291920 + .203440i
0.2	−.005120 + .053760i	.035985 + .052395i	.083840 + .050880i	.146995 + .054165i	.230400 + .067200i	.336005 + .093135i
0.1	−.002835 + .016605i	.029320 + .011640i	.074975 − .004825i	.143280 − .023040i	.237985 − .033855i	.358040 − .030520i
0.0	−.000000 + .000000i	.008155 − .017415i	.042560 − .058880i	.112365 − .108045i	.219520 − .149760i	.359375 − .171875i

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	−.990720 + .192960i	−1.332065 + .563745i	−1.653760 + 1.134080i	−1.899855 + 1.937115i	−2.000000 + 3.000000i
0.9	−.701685 + .453555i	−.882080 + .811440i	−1.013875 + 1.322125i	−1.046520 + 2.005560i	−.916865 + 2.875095i
0.8	−.420800 + .585600i	−.478695 + .905535i	−.471040 + 1.336320i	−.353285 + 1.886505i	−.070080 + 2.557440i
0.7	−.169515 + .620145i	−.139160 + .887880i	−.036505 + 1.229915i	.176400 + 1.645200i	.546505 + 2.124885i
0.6	.040320 + .584640i	.129475 + .794925i	.289280 + 1.048960i	.550485 + 1.337895i	.951040 + 1.644480i
0.5	.204655 + .502335i	.328560 + .657120i	.514865 + .831705i	.786520 + 1.011240i	1.171875 + 1.171875i
0.4	.325440 + .391680i	.466045 + .498315i	.656000 + .608000i	.909855 + .701685i	1.245760 + .750720i
0.3	.408825 + .265725i	.554680 + .335160i	.733835 + .398695i	.951840 + .434880i	1.216045 + .412065i
0.2	.463360 + .131520i	.610215 + .176505i	.773120 + .215040i	.948025 + .225075i	1.130880 + .173760i
0.1	.498195 − .010485i	.649600 + .022800i	.800405 + .058085i	.936360 + .074520i	1.041415 + .039855i
0.0	.521280 − .167040i	.689185 − .134505i	.842240 − .081920i	.955395 − .027135i	1.000000 + .000000i

TABLES OF COEFFICIENTS $A_k^{(n)}(P)$ AND $B_k^{(n)}(P)$

Three-Point: $B_q^{(3)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	.00000 + .00000i	−.01781 + .00370i	−.06112 + .02720i	−.11223 + .06370i	−.15104 + .17920i	−.15625 + .31250i
0.9	.01620 + .00171i	.00324 + .02916i	−.02022 + .06371i	−.04308 + .11556i	−.05184 + .18891i	−.03180 + .28196i
0.8	.05120 + .01152i	.03539 + .04652i	.02048 + .08192i	.01217 + .12492i	.01856 + .17792i	.04895 + .23852i
0.7	.08820 + .03213i	.06724 + .06268i	.05498 + .09173i	.05292 + .12348i	.06496 + .15853i	.09620 + .19388i
0.6	.11520 + .06144i	.09099 + .07974i	.07968 + .09824i	.07977 + .11814i	.09216 + .13824i	.11895 + .15494i
0.5	.12500 + .09375i	.10244 + .09620i	.09338 + .10295i	.09452 + .11220i	.10496 + .12095i	.12500 + .12500i
0.4	.11520 + .12096i	.10099 + .10816i	.09728 + .10496i	.10017 + .10656i	.10816 + .10816i	.12095 + .10496i
0.3	.08820 + .13377i	.08964 + .11052i	.09498 + .10217i	.10092 + .10092i	.10656 + .10017i	.11220 + .09452i
0.2	.05120 + .12288i	.07499 + .09818i	.09248 + .09248i	.10217 + .09498i	.10496 + .09728i	.10295 + .09338i
0.1	.01620 + .08019i	.06724 + .06724i	.09818 + .07499i	.11052 + .08964i	.10816 + .10099i	.09620 + .10244i
0.0	.00000 + .00000i	.08019 + .01620i	.12288 + .05120i	.13377 + .08820i	.12096 + .11520i	.09375 + .12500i

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	−.10656 + .47520i	.01813 + .65170i	.23552 + .81920i	.55971 + .94770i	1.00000 + 1.00000i
0.9	.03174 + .38691i	.15228 + .48996i	.34092 + .57131i	.60516 + .60516i	.94770 + .55971i
0.8	.11264 + .29952i	.21773 + .34892i	.36992 + .36992i	.57131 + .34092i	.81920 + .23552i
0.7	.15174 + .22293i	.23548 + .23548i	.34892 + .21773i	.48996 + .15228i	.65170 + .01813i
0.6	.16224 + .16224i	.22293 + .15174i	.29952 + .11264i	.38691 + .03174i	.47520 − .10656i
0.5	.15494 + .11895i	.19388 + .09620i	.23852 + .04895i	.28196 − .03180i	.31250 − .15625i
0.4	.13824 + .09216i	.15853 + .06496i	.17792 + .01856i	.18891 − .05184i	.17920 − .15104i
0.3	.11814 + .07977i	.12348 + .05292i	.12492 + .01217i	.11556 − .04308i	.08370 − .11223i
0.2	.09824 + .07968i	.09173 + .05498i	.08192 + .02048i	.06371 − .02022i	.02720 − .06112i
0.1	.07974 + .09099i	.06268 + .06724i	.04652 + .03539i	.02916 + .00324i	.00370 − .01781i
0.0	.06144 + .11520i	.03213 + .08820i	.01152 + .05120i	.00171 + .01620i	.00000 − .00000i

Three-Point: $B_1^{(3)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	−.000000 − .000000i	.005850 + .003455i	.025600 + .007360i	.059850 + .001665i	.105600 − .023680i	.156250 − .078125i
0.9	−.003645 − .004050i	.005580 − .008820i	.020555 − .015240i	.041580 − .030060i	.065955 − .060030i	.087980 − .111300i
0.8	−.010240 − .012800i	−.000940 − .019545i	.008960 − .029440i	.019460 − .046535i	.028160 − .074880i	.030260 − .117925i
0.7	−.015435 − .022050i	−.008460 − .027220i	−.003635 − .036440i	−.001260 − .051660i	−.003435 − .074830i	−.014060 − .107300i
0.6	−.017280 − .028800i	−.014130 − .030945i	−.014080 − .037440i	−.017730 − .048735i	−.026880 − .065280i	−.044530 − .086925i
0.5	−.015625 − .031250i	−.016900 − .030420i	−.021025 − .033640i	−.028900 − .040460i	−.042025 − .050430i	−.062500 − .062500i
0.4	−.011520 − .028800i	−.016920 − .025945i	−.024320 − .026240i	−.034920 − .028935i	−.049920 − .033280i	−.070520 − .037925i
0.3	−.006615 − .022050i	−.014940 − .018420i	−.024415 − .016440i	−.036540 − .015660i	−.052215 − .015630i	−.071740 − .015300i
0.2	−.002560 − .012800i	−.011710 − .009345i	−.021760 − .005440i	−.034510 − .001535i	−.050560 + .001920i	−.069310 + .005075i
0.1	−.000405 − .004050i	−.007380 − .000820i	−.016205 + .005560i	−.028980 + .013140i	−.046005 + .019970i	−.065780 + .024700i
0.0	−.000000 − .000000i	−.000900 + .004455i	−.006400 + .015360i	−.018900 + .028665i	−.038400 + .040320i	−.062500 + .046875i

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	.201600 − .169920i	.227850 − .305515i	.217600 − .488960i	.149850 − .721305i	.000000 − 1.000000i
0.9	.098955 − .188820i	.087180 − .295740i	.037955 − .432810i	−.066420 − .597780i	−.246645 − .784800i
0.8	.018560 − .177920i	−.016540 − .255315i	−.087040 − .348160i	−.207340 − .451505i	−.394240 − .556800i
0.7	−.038835 − .149220i	−.085260 − .198940i	−.162635 − .252410i	−.282060 − .302580i	−.456435 − .338800i
0.6	−.074880 − .112320i	−.123330 − .138315i	−.196480 − .159360i	−.302130 − .166905i	−.449280 − .148800i
0.5	−.093025 − .074420i	−.136900 − .082140i	−.198025 − .079210i	−.280900 − .056180i	−.390625 − .000000i
0.4	−.097920 − .040320i	−.133320 − .036115i	−.177920 − .018560i	−.232920 + .022095i	−.299520 + .099200i
0.3	−.094815 − .012420i	−.120540 − .002940i	−.147415 + .019590i	−.173340 + .064620i	−.195615 + .145200i
0.2	−.088960 + .009280i	−.106510 + .017685i	−.117760 + .035840i	−.117310 + .072295i	−.098560 + .139200i
0.1	−.085005 + .027180i	−.098580 + .029060i	−.099605 + .034390i	−.079380 + .050220i	−.027405 + .087200i
0.0	−.086400 + .046080i	−.102900 + .037485i	−.102400 + .023040i	−.072900 + .007695i	−.000000 + .000000i

TABLES OF COEFFICIENTS $A_k^{(n)}(P)$ AND $B_k^{(n)}(P)$

Four-Point: $A_k^{(4)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	.00000000 − .00000000i	.03342465 + .01637415i	.10526720 + .02795520i	.17487855 + .00972405i	.21288960 − .03363840i	.20703125 − .08203125i
0.9	−.02486465 − .03368385i	.04276800 − .04484160i	.10623345 − .03779335i	.15610960 − .03744720i	.18235035 − .04799925i	.18089760 − .06192160i
0.8	−.04734720 − .13383680i	.05435855 − .11928735i	.12083200 − .09338880i	.16210005 − .07178825i	.18095360 − .05844480i	.17959555 − .05008155i
0.7	−.01179855 − .27339795i	.09825840 − .20247120i	.15955595 − .14271825i	.19051200 − .09666720i	.20023245 − .06346215i	.19443920 − .03953520i
0.6	.10455040 − .40688640i	.18787365 − .27398925i	.22717440 − .17774080i	.23967955 − .10822815i	.23500800 − .05898240i	.21962265 − .02462665i
0.5	.29296875 − .48828125i	.31910240 − .31011360i	.32040445 − .18653955i	.30556080 − .09966320i	.28037735 − .03962865i	.25000000 − .00000000i
0.4	.51989760 − .48591360i	.47401875 − .29414835i	.42936320 − .15989760i	.38171385 − .06476445i	.33096960 − .00000000i	.28037735 + .03962865i
0.3	.73815805 − .39361455i	.62833680 − .22379760i	.54204375 − .09564205i	.46154080 − .00000000i	.38171385 + .06476445i	.30556080 + .09966320i
0.2	.90193920 − .23736320i	.76189865 − .11478945i	.65105920 − .00000000i	.54204375 + .09564205i	.42936320 + .15989760i	.32040445 + .18653955i
0.1	.98481015 − .07667865i	.87143040 − .00000000i	.76189865 + .11478945i	.62833680 + .22379760i	.47401875 + .29414835i	.31910240 + .31011360i
0.0	1.00000000 − .00000000i	.98481015 + .07667865i	.90193920 + .23736320i	.73815805 + .39361455i	.51989760 + .48591360i	.29296875 + .48828125i

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	.16266240 − .11266560i	.09876195 − .11049255i	.04014080 − .07557120i	.00662985 − .02662065i	.00000000 + .00000000i
0.9	.15575725 − .06815835i	.11729520 − .05877360i	.07750935 − .03329545i	.04303360 − .00000000i	.00662985 + .02662065i
0.8	.16250880 − .04060160i	.13630025 − .02471205i	.10727680 − .00000000i	.07750935 + .03329545i	.04014080 + .07557120i
0.7	.17837415 − .01999845i	.15743520 − .00000000i	.13630025 + .02471205i	.11729520 + .05877360i	.09876195 + .11049255i
0.6	.19901440 − .00000000i	.17837415 + .01999845i	.16250880 + .04060160i	.15575725 + .06815835i	.16266240 + .11266560i
0.5	.21962265 + .02462665i	.19443920 + .03953520i	.17959555 + .05008155i	.18089760 + .06192160i	.20703125 + .08203125i
0.4	.23500800 + .05898240i	.20023245 + .06346215i	.18095360 + .05844480i	.18235035 + .04799925i	.21288960 + .03363840i
0.3	.23967955 + .10822815i	.19051200 + .09666720i	.16210005 + .07178825i	.15610960 + .03744720i	.17487855 − .00972405i
0.2	.22717440 + .17774080i	.15955595 + .14271825i	.12083200 + .09338880i	.10623345 + .03779335i	.10526720 − .02795520i
0.1	.18787365 + .27398925i	.09825840 + .20247120i	.05435855 + .11928735i	.04276800 + .04484160i	.03342465 − .01637415i
0.0	.10455040 + .40688640i	−.01179855 + .27339795i	−.04734720 + .13383680i	−.02486465 + .03368385i	.00000000 − .00000000i

Four-Point: $B_0^{(4)}(P)$

q	p = 0.0	0.1	0.2	0.3	0.4	0.5
1.0	.00000000 − .00000000i	.00149850 + .00721305i	.00870400 + .01955840i	.02050650 + .02749635i	.03225600 + .02718720i	.03906250 + .01953125i
0.9	.00246645 − .00784800i	.01195560 + .00132840i	.01845105 + .01146610i	.02471880 + .01842840i	.02994885 + .02040660i	.03224520 + .01791400i
0.8	.01576960 − .02227200i	.02428040 − .00525155i	.02785280 + .00696320i	.02981080 + .01475055i	.03069440 + .01838080i	.02993960 + .01871225i
0.7	.04107915 − .03049200i	.04071960 − .00728280i	.03842095 + .00689570i	.03580920 + .01534680i	.03309435 + .01976580i	.02994040 + .02138600i
0.6	.07188480 − .02380800i	.05867190 − .00086535i	.04907520 + .01231360i	.04182870 + .01991715i	.03594240 + .02396160i	.03076230 + .02563525i
0.5	.09765625 − .00000000i	.07303400 + .01460680i	.05742725 + .02297090i	.04654600 + .02792760i	.03814025 + .03051220i	.03125000 + .03125000i
0.4	.10782720 + .03571200i	.07887060 + .03568365i	.06138880 + .03676160i	.04899780 + .03824415i	.03893760 + .03893760i	.03051220 + .03814025i
0.3	.09585135 + .07114800i	.07415640 + .05528520i	.06085155 + .05009170i	.04945080 + .04945080i	.03824415 + .04899780i	.02792760 + .04654600i
0.2	.06307840 + .08908800i	.06233810 + .06431545i	.05918720 + .05918720i	.05009170 + .06085155i	.03676160 + .06138880i	.02297090 + .05742725i
0.1	.02219805 + .07063200i	.05446440 + .05446440i	.06431545 + .06233810i	.05528520 + .07415640i	.03568365 + .07887060i	.01460680 + .07303400i
0.0	.00000000 + .00000000i	.07063200 + .02219805i	.08908800 + .06307840i	.07114800 + .09585135i	.03571200 + .10782720i	.00000000 + .09765625i

q	p = 0.6	0.7	0.8	0.9	1.0
1.0	.03801600 + .00852480i	.02932650 − .00081585i	.01638400 − .00471040i	.00473850 − .00279855i	− .00000000 + .00000000i
0.9	.03028785 + .01309590i	.02416680 + .00860760i	.01538805 + .00631240i	.00605160 + .00605160i	− .00279855 + .00473850i
0.8	.02698240 + .01720320i	.02178680 + .01549195i	.01479680 + .01479680i	.00631240 + .01538805i	− .00471040 + .01638400i
0.7	.02601135 + .02144070i	.02119320 + .02119320i	.01549195 + .02178680i	.00860760 + .02416680i	− .00081585 + .02932650i
0.6	.02595840 + .02595840i	.02144070 + .02601135i	.01720320 + .02698240i	.01309590 + .03028785i	.00852480 + .03801600i
0.5	.02563525 + .03076230i	.02138600 + .02994040i	.01871225 + .02993960i	.01791400 + .03224520i	.01953125 + .03906250i
0.4	.02396160 + .03594240i	.01976580 + .03309435i	.01838080 + .03069440i	.02040660 + .02994885i	.02718720 + .03225600i
0.3	.01991715 + .04182870i	.01534680 + .03580920i	.01475055 + .02981080i	.01842840 + .02471880i	.02749635 + .02050650i
0.2	.01231360 + .04907520i	.00689570 + .03842095i	.00696320 + .02785280i	.01146610 + .01845105i	.01955840 + .00870400i
0.1	− .00086535 + .05867190i	− .00728280 + .04071960i	− .00525155 + .02428040i	.00132840 + .01195560i	.00721305 + .00149850i
0.0	− .02380800 + .07188480i	− .03049200 + .04107915i	− .02227200 + .01576960i	− .00784800 + .00246645i	.00000000 + .00000000i

$$A_i^{(3)}(p+iq) = \overline{A_i^{(3)}(q+ip)}, \tag{6}$$

$$B_i^{(3)}(p+iq) = i\,\overline{B_i^{(3)}(q+ip)}; \tag{6'}$$

$$A_1^{(4)}(p+iq) = \overline{A_0^{(4)}(1-p+iq)}, \tag{7}$$

$$B_1^{(4)}(p+iq) = -\overline{B_0^{(4)}(1-p+iq)}, \tag{7'}$$

$$A_i^{(4)}(p+iq) = \overline{A_0^{(4)}(p+i\,(1-q))}, \tag{8}$$

$$B_i^{(4)}(p+iq) = \overline{B_0^{(4)}(p+i\,(1-q))}, \tag{8'}$$

$$A_{1+i}^{(4)}(p+iq) = A_0^{(4)}(1-p+i\,(1-q)), \tag{9}$$

$$B_{1+i}^{(4)}(p+iq) = -B_0^{(4)}(1-p+i\,(1-q)). \tag{9'}$$

The tables of $A_k^{(n)}(P)$ and $B_k^{(n)}(P)$ were calculated by first squaring the ordinary Lagrangian coefficients $L_k^{(n)}(P)$ which have already been published [4], and then multiplying by the appropriate linear functions of P. The results were checked in a preliminary manner by evaluating both $\sum_k A_k^{(n)}(P) = 1$ and $(3+2i)+.5P = \sum_k \{3+2i+.5k\}A_k^{(n)}(P) + .5\sum_k B_k^{(n)}(P)$ for all values of P. The following final functional check (which was designed also to verify the highest degree accuracy that was attainable with the set of coefficients corresponding to each n) was then performed for all values of P:

$$(1+3i+P)^{2n-1} = \sum_k \{1+3i+k\}^{2n-1}A_k^{(n)}(P) + (2n-1)\sum_k \{1+3i+k\}^{2n-2}B_k^{(n)}(P).$$

For osculatory interpolation of degree higher than seven, or for P given finer than 0.1 in p or q, when these present tables might not be adequate, one is referred to an article [9] by the author which describes a longer method, but which furnishes up to 13th degree accuracy and which can be used for P given to any number of decimals and even outside the range $0 \leqq p \leqq 1$, $0 \leqq q \leqq 1$.

The author wishes to express his appreciation to Miss Isabelle Arsham for her assistance in the computation and checking of these tables.

REFERENCES

[1] T. Fort, *Finite Differences*, Clarendon Press, Oxford, 1948, pp. 85–88.

[2] J. F. Steffensen, *Interpolation*, Williams and Wilkins, Baltimore, 1927, pp. 33–34.

[3] P. Montel, "Sur quelques propriétés des différences divisées," *Journal de Mathématique Pures et Appliquées*, 9 ser., vol. 16, 1937, pp. 219–231.

[4] A. N. Lowan and H. E. Salzer, "Coefficients for Interpolation within a Square Grid in the Complex Plane," *Journal of Math. and Phys.*, vol. 23, no. 3, 1944, pp. 156–166.

[5] H. E. Salzer, "Coefficients for Complex Quartic, Quintic, and Sextic Interpolation within a Square Grid," *Jour. of Math. and Phys.*, vol. 27, no. 2, 1948, pp. 136–156.

[6] National Bureau of Standards, *Tables of the Error Function and Its Derivative*, forthcoming.

[7] National Bureau of Standards, *Tables of Exponential Integrals for Complex Arguments*, forthcoming.

[8] HARVARD COMPUTATION LABORATORY, *Tables of the Modified Hankel Functions of Order One-Third and of their Derivatives*, Harvard University Press, Cambridge, Mass., 1945.

[9] H. E. SALZER, "Osculatory Interpolation in the Complex Plane," *Jour. Res. Nat. Bur. Stands*, vol. 54, 1955, pp. 263–266.

CONVAIR ASTRONAUTICS
SAN DIEGO, CALIFORNIA
(This work was done while the author was at the Diamond Ordnance Fuse Laboratories, Washington 25, D.C.)

(Received May 26, 1955)

A POWER SERIES SOLUTION OF THE NONLINEAR EQUATIONS FOR AXI-SYMMETRICAL BENDING OF SHALLOW SPHERICAL SHELLS*

By Roger M. Simons

1. Introduction. This paper is concerned with the stresses and deformations of thin shallow segments of spherical shells subjected to uniform normal pressure. Two important types of edge restraints will be considered, namely a simply supported edge and a clamped edge.

It will be seen that there are two essential parameters for this problem: μ, a geometrical parameter, which is zero for a flat plate and large compared with one for a sufficiently thin shell, and γ, a load parameter, which is a factor of the nonlinear terms in the differential equations of the problem. The linearized differential equations, for small deflections, can be solved explicitly in terms of Bessel functions [1].† It is the purpose of the present paper to obtain a solution of the problem for small finite deflections in which case the differential equations of the problem are nonlinear [3]. This solution will be obtained in the form of a power series in powers of a coordinate. A solution of this type has been obtained by Way [4] for the special case of a flat circular plate. A serious shortcoming of the power series solution obtained by Way is that the boundary conditions satisfied by the solution are known only after the solution is obtained. The only way a prescribed set of boundary conditions can be satisfied is to obtain several power series solutions (by choosing the two leading coefficients in the power series differently each time) and then using interpolation to find the appropriate choices of the leading coefficients to satisfy the prescribed boundary conditions. The interpolation is difficult because there are two independent constants to determine. A principal contribution of this paper is to give a systematic procedure for obtaining the required leading coefficients for the power series solution in order to satisfy the boundary conditions. The technique is given for a shallow shell and contains as a special case the corresponding results for a flat plate.

2. Notation and basic equations. The following notation will be used in this paper:

r, θ:	polar coordinates
ξ:	angle subtended at the center of the sphere between apex and variable point
α:	half the opening angle of the shell, that is the value of ξ at the edge of the shell

* Portion of a Ph.D. Thesis accepted by the Massachusetts Institute of Technology in June 1955. The author wishes to express thanks to his thesis advisor, Professor Eric Reissner, for guidance and advice; and to the Office of Naval Research of the U.S. Navy for financial support under Contract N5-ori-7834 with the Massachusetts Institute of Technology.

† Numbers in square brackets refer to the references at the end of this paper.

E: Young's modulus of elasticity
ν: Poisson's ratio
a: radius of the middle surface of the shell
$R = a\alpha$: half-span of the shallow shell
h: thickness of the shell, assumed constant
$D = Eh^3/m^4$: flexural rigidity
p: uniform normal pressure
$\mu = m\sqrt{\frac{R}{h}}\alpha$, $m = [12(1-\nu^2)]^{\frac{1}{4}}$, $\gamma = \frac{m^6}{4}\frac{pR^4}{Eh^4}$
N_ξ, N_θ, Q: stress resultants in meridional and circumferential directions and shear stress resultant, respectively
M_ξ, M_θ: stress couples in meridional and circumferential directions, respectively
H: horizontal component of stress resultants N_ξ and Q
u, w: horizontal and vertical components of displacement
β: negative of the angle of rotation of the tangent to a meridian
$\Psi = rH$: a stress function

The differential equations for small finite deflections of a thin shallow spherical shell of uniform thickness subjected to a uniform normal pressure can be obtained by suitably specializing equations (60) and (61) in [3] as follows:

$$\begin{aligned} \beta'' + \frac{1}{\xi}\beta' - \frac{1}{\xi^2}\beta &= -\frac{a}{D}\left(\Psi - \tfrac{1}{2}p a^2\xi - \frac{1}{\xi}\beta\Psi\right), \\ \Psi'' + \frac{1}{\xi}\Psi' - \frac{1}{\xi^2}\Psi &= aEh\left(\beta - \frac{1}{2\xi}\beta^2\right). \end{aligned} \tag{1}$$

In these equations primes indicate differentiation with respect to ξ.

Stress resultants, couples, and displacements are obtained from equations (62) and (63) of [3] similarly as follows in terms of β and Ψ.

$$\begin{aligned} aN_\xi &= \frac{\Psi}{\xi}, \qquad aN_\theta = \Psi', \qquad aQ = -\Psi + \tfrac{1}{2}pa^2\xi + \frac{\beta\Psi}{\xi}, \\ aM_\xi &= D\left(\beta' + \nu\frac{\beta}{\xi}\right), \qquad aM_\theta = D\left(\nu\beta' + \frac{\beta}{\xi}\right), \\ Ehu &= \left(\Psi' - \nu\frac{\Psi}{\xi}\right)\xi, \qquad w = -a\int \beta\, d\xi. \end{aligned} \tag{2}$$

At a simply supported edge, the meridional bending couple M_ξ and the radial displacement u vanish. In terms of β and Ψ, the boundary conditions for a simply supported edge are

$$\alpha\beta'(\alpha) + \nu\beta(\alpha) = 0, \qquad \alpha\Psi'(\alpha) - \nu\Psi(\alpha) = 0. \tag{3}$$

At a clamped edge, the rotation of the meridian β and the radial displacement u vanish. In terms of β and Ψ, the boundary conditions for a clamped edge are

$$\beta(\alpha) = 0, \qquad \alpha\Psi'(\alpha) - \nu\Psi(\alpha) = 0. \tag{4}$$

Observing that $\xi = 0$ is a singular point of the system of differential equations (1), we require for a shell without a central hole that β and Ψ be finite at the apex; hence, two additional boundary conditions are

$$\beta(0) < \infty, \qquad \Psi(0) < \infty. \tag{5}$$

It will be observed that the equations applicable to the special case of a flat plate cannot be obtained directly from (1), since in the transition from a shell to a plate, a becomes infinite and α tends to zero, so that the range of $\xi(0 \leqq \xi \leqq \alpha)$ shrinks to zero. Motivated by the desire to have a solution of the shallow shell problem which contains as a special case the solution of the corresponding flat plate problem, and also in order to non-dimensionalize the problem, we introduce new variables defined by the following relations:

$$\xi = \alpha x, \qquad \beta = \frac{m^4}{4}\frac{\rho R^2}{Eh^2}F, \qquad \Psi = \frac{m^2}{4}\frac{\rho R^2}{h}G, \tag{6}$$

where $R = a\alpha$. In terms of the variables F, G, and x, and the parameters γ and μ, the differential equations (1) become:

$$\begin{aligned} F'' + \frac{1}{x}F' - \frac{1}{x^2}F &= -\mu^2 G + 2x + \gamma\frac{FG}{x}, \\ G'' + \frac{1}{x}G' - \frac{1}{x^2}G &= \mu^2 F - \gamma\frac{F^2}{2x}, \end{aligned} \tag{7}$$

where primes denote differentiation with respect to x. The differential equations in this form can be specialized to the case of a flat plate simply by setting $\mu = 0$. The quantity $\gamma = \frac{1}{4}m^6\rho R^4/Eh^4$ is unaffected by the transition from shell to plate.

The boundary conditions (3), (4), and (5) become in terms of the new variables

$$F'(1) + \nu F(1) = 0, \qquad G'(1) - \nu G(1) = 0, \tag{8}$$

$$F(1) = 0, \qquad G'(1) - \nu G(1) = 0, \tag{9}$$

$$F(0) < \infty, \qquad G(0) < \infty. \tag{10}$$

The stress resultants, couples, and displacements (2) become

$$\begin{aligned} N_\xi &= \frac{m^2}{4}\frac{\rho R^2}{h}\frac{G}{x}, \qquad N_\theta = \frac{m^2}{4}\frac{\rho R^2}{h}G', \\ Q &= \tfrac{1}{4}\rho R\left(-\mu^2 G + 2x + \gamma\frac{FG}{x}\right), \\ M_\xi &= \tfrac{1}{4}\rho R^2\left(F' + \nu\frac{F}{x}\right), \qquad M_\theta = \tfrac{1}{4}\rho R^2\left(\nu F' + \frac{F}{x}\right), \\ \frac{u}{h} &= \frac{1}{m^4}\frac{\gamma h}{R}\left(G' - \nu\frac{G}{x}\right)x, \qquad \frac{w}{h} = -\frac{\gamma}{m^2}\int F\,dx. \end{aligned} \tag{11}$$

We shall be concerned in what follows with the system of equations (7) for $0 \leqq x \leqq 1$, $\mu \geqq 0$, and the boundary conditions (8) and (10) or (9) and (10).

The relation $\gamma > 0$ indicates outward pressure, and $\gamma < 0$ indicates inward pressure, in which case the possibility of the shell becoming unstable for sufficiently large $-\gamma$ will be considered.

3. Power series solution. We attempt to find solutions of (7) in the form

$$F = x^s \sum_{n=0}^{\infty} f_n x^n, \qquad G = x^s \sum_{n=0}^{\infty} g_n x^n, \tag{12}$$

where s is a constant to be determined, and not both f_0 and g_0 are zero. Substituting (12) into (7), we find four independent solutions: two with $s = 1$ and two with $s = -1$. In order that the regularity condition (10) be satisfied, we retain only those solutions for which $s = 1$. Further, one readily finds that the coefficients of all even powers of x must be zero. For simplicity of notation, it will be understood in the future that all indices of summation run over the odd integers only, and we shall write in place of (12)

$$F = \sum F_n x^n, \qquad G = \sum G_n x^n. \tag{13}$$

Substituting the series (13) into equations (7) and equating coefficients of like powers of x, one finds recursion formulas for the coefficients of the series as follows:

$$\begin{aligned}
F_3 &= \tfrac{1}{8}(2 - \mu^2 G_1 + \gamma F_1 G_1), \\
F_n &= \frac{1}{n^2 - 1}\left(-\mu^2 G_{n-2} + \gamma \sum_{j=1}^{n-2} F_j G_{n-j-1}\right), \qquad n = 5, 7, \cdots, \\
G_n &= \frac{1}{n^2 - 1}\left(\mu^2 F_{n-2} - \frac{\gamma}{2} \sum_{j=1}^{n-2} F_j F_{n-j-1}\right), \qquad n = 3, 5, \cdots.
\end{aligned} \tag{14}$$

In terms of the power series solution, the expressions (11) for the stress resultants, couples, and displacements become

$$\begin{aligned}
N_\xi &= \frac{m^2}{4}\frac{\rho R^2}{h} \sum G_n x^{n-1}, \qquad N_\theta = \frac{m^2}{4}\frac{\rho R^2}{h} \sum n G_n x^{n-1}, \\
Q &= \tfrac{1}{4}\rho R \sum (n^2 - 1) F_n x^{n-2}, \\
M_\xi &= \tfrac{1}{4}\rho R^2 \sum (n + \nu) F_n x^{n-1}, \qquad M_\theta = \tfrac{1}{4}\rho R^2 \sum (\nu n + 1) F_n x^{n-1}, \\
\frac{u}{h} &= \frac{1}{m^4}\frac{\gamma h}{R} \sum (n - \nu) G_n x^n, \qquad \frac{w}{h} = \frac{1}{m^2}\gamma \sum \left[\frac{F_n}{n+1}(1 - x^{n+1})\right].
\end{aligned} \tag{15}$$

Once F_1 and G_1 are determined, all succeeding coefficients can be computed from the recursion formulas (14), and the corresponding expressions for the stress resultants, couples, and displacements can be computed from (15). The problem, then, is to determine F_1 and G_1 in such a way that the boundary conditions are satisfied.

4. Initial estimates of the leading coefficients. We shall consider first the case of a simply supported edge, the boundary conditions for which are (8).

If the problem were linear, that is if γ were zero, the solution could be expressed in closed form in terms of Kelvin functions. In particular, this solution can be expanded in a power series of odd powers of x.

As one means of obtaining an initial estimate of F_1 and G_1 (denoted by $F_1^{(0)}$ and $G_1^{(0)}$, respectively), we choose the leading coefficients in the series solution of the linearized problem:

$$F_1^{(0)} = \frac{1-\nu}{\sqrt{2}\mu}\,\frac{\mu(ber_1'\mu + bei_1'\mu) + \nu(ber_1\mu + bei_1\mu)}{\mu^2(ber_1'^2\mu + bei_1'^2\mu) - \nu^2(ber_1^2\mu + bei_1^2\mu)},$$
$$G_1^{(0)} = \frac{1-\nu}{\sqrt{2}\mu}\,\frac{\mu(ber_1'\mu - bei_1'\mu) + \nu(ber_1\mu - bei_1\mu)}{\mu^2(ber_1'^2\mu + bei_1'^2\mu) - \nu^2(ber_1^2\mu + bei_1^2\mu)} + \frac{2}{\mu^2}. \tag{16}$$

The Kelvin functions in (16) can be defined as the real and imaginary parts of a Bessel function with complex argument in the form:

$$J_1(i^{\frac{3}{4}}\mu) \equiv ber_1\,\mu + i bei_1\,\mu.$$

It is reasonable to expect that the above initial estimate of F_1 and G_1 is fairly accurate, provided γ is not too large.

A limiting case of the linear shell problem is the linear plate problem for which $\mu = 0$ as well as $\gamma = 0$. The solution of the linear plate problem is

$$F = -\frac{3+\nu}{4(1+\nu)}\,x + \tfrac{1}{4}x^3, \qquad G = 0. \tag{17}$$

Hence, provided μ is small, much simpler initial estimates of the leading coefficients of the series solution than are provided by (16) are given by

$$F_1^{(0)} = -\frac{3+\nu}{4(1+\nu)}, \qquad G_1^{(0)} = 0. \tag{18}$$

For $\nu = \frac{1}{3}$, (18) becomes $F_1^{(0)} = -.625$, $G_1^{(0)} = 0$. For the same value of ν, the initial estimates furnished by (16) for representative values of μ^2 are given in Table 1.

The foregoing estimates, (16) and (18), of the leading coefficients in the series solution can be improved by taking account of the effect of non-linearity. F and G have power series expansions in odd powers of x, and for the special case of linear plate theory, these functions are given by (17). Hence, we expect for moderate values of γ and μ, the polynomials

$$F_k = Ax + Bx^3, \qquad G_k = Cx + Dx^3, \tag{19}$$

TABLE 1

Initial estimate of F_1 and G_1 given by linear theory for several values of μ^2

	μ^2				
	0	1	3	5	10
$-F_1^{(0)}$	.625	.568	.323	.167	.0409
$G_1^{(0)}$	.000	.219	.382	.346	.216

to provide—for appropriate values of A, B, C, D—the major contribution of the series solution.

In order to determine appropriate values of A, B, C, and D, we shall require that F_k and G_k satisfy the boundary conditions (8), and also that F_k and G_k satisfy the differential equation (7) in the mean over the interval $0 \leqq x \leqq 1$, in the sense that

$$\begin{aligned} \int_0^1 \left(F_k'' + \frac{1}{x} F_k' - \frac{1}{x^2} F_k + \mu^2 G_k - 2x - \gamma \frac{F_k G_k}{x} \right) dx = 0, \\ \int_0^1 \left(G_k'' + \frac{1}{x} G_k' - \frac{1}{x^2} G_k - \mu^2 F_k + \gamma \frac{F_k^2}{2x} \right) dx = 0. \end{aligned} \tag{20}$$

These requirements on A, B, C, and D determine A as a root of the equation

$$C_1 \gamma^2 A^3 + C_2 \gamma \mu^2 A^2 + (C_3 \mu^4 + C_4) A + C_5 = 0, \tag{21}$$

and determine C as

$$C = C_6 \mu^2 A + C_7 \gamma A^2, \tag{22}$$

where

$$\begin{aligned} &C_1 = 2(19 - \nu^2)(19 + 8\nu + \nu^2), \quad C_2 = -3(3 + \nu)(285 + 59\nu - 13\nu^2 - 3\nu^3), \\ &C_3 = 9(3 + \nu)^2(25 - \nu^2), \quad C_4 = 2304(3 + \nu)^2(1 - \nu^2), \\ &C_5 = 576(3 + \nu)^3(1 - \nu), \\ &C_6 = -\frac{(3 - \nu)(5 + \nu)}{16(1 - \nu)(3 + \nu)}, \qquad C_7 = \frac{(3 - \nu)(19 + 8\nu + \nu^2)}{48(1 - \nu)(3 + \nu)^2}. \end{aligned} \tag{23}$$

As a third means of determining an initial estimate of the leading coefficients in the power series solution, we take

$$F_1^{(0)} = A, \qquad G_1^{(0)} = C, \tag{24}$$

as determined by (21) and (22).

Values of A and C which have been computed for $\nu = \frac{1}{3}$ and for several values of γ and μ^2, are listed in Tables 2 and 3.

To obtain an idea of the accuracy of the averaging method for obtaining initial estimates of F_1 and G_1, we compare the estimates furnished by the averaging method for $\gamma = 0$ with the exact values of F_1 and G_1 from the linear theory. Specifically, comparison of Tables 2 and 3 with (18) for $\nu = \frac{1}{3}$ indicates that the averaging technique yields the correct values of F_1 and G_1 for linearized plate theory, that is the case $\gamma = \mu = 0$. Comparison of Tables 2 and 3 for $\gamma = 0$ with Table 1 indicates that the averaging technique yields initial estimates of F_1 and G_1 which are, in the range of μ-values which is considered, in close agreement with the actual values of F_1 and G_1 from the linearized shell theory.

5. Successive refinements of the values of the leading coefficients. We now give a method by which the coefficients F_1 and G_1 can be made successively more

TABLE 2

Initial estimate of $-F_1$ given by averaging method for several values of γ and μ^2, $\nu = \frac{1}{2}$

γ	μ^2				
	0	1	3	5	10
−10	.398	.480	.671	.309	.0558
−5	.507	.603	.530	.202	.0540
−3	.566	.633	.402	.185	.0533
−1	.617	.600	.337	.173	.0527
0	.625	.563	.315	.167	.0524
1	.617	.525	.297	.163	.0521
3	.566	.459	.268	.154	.0515
5	.507	.408	.247	.147	.0509
10	.398	.325	.210	.133	.0496

TABLE 3

Initial estimate of G_1 given by averaging method for several values of γ and μ^2, $\nu = \frac{1}{2}$

γ	μ^2				
	0	1	3	5	10
−10	−.259	−.184	.0698	.462	.218
−5	−.210	−.0557	.406	.371	.214
−3	−.157	.0569	.403	.353	.212
−1	−.0622	.181	.386	.341	.210
0	.0000	.225	.378	.334	.210
1	.0622	.255	.371	.330	.209
3	.157	.287	.357	.320	.207
5	.210	.299	.346	.312	.206
10	.259	.302	.324	.295	.202

accurate. Suppose that the m^{th} approximation of F_1 and G_1, and therefore of F_n and G_n, have been made. Denote the m^{th} approximation of F_n and G_n by $F_n^{(m)}$ and $G_n^{(m)}$, respectively. Section 4 deals with the zero$^{\text{th}}$ approximation ($m = 0$). We now find the $(m+1)^{\text{th}}$ approximation.

Let

$$F_n^{(m+1)} = F_n^{(m)} + \Delta F_n^{(m)}, \qquad G_n^{(m+1)} = G_n^{(m)} + \Delta G_n^{(m)}. \tag{25}$$

First, we require that $F_n^{(m+1)}$ and $G_n^{(m+1)}$ satisfy the recursion formulas (14), except for terms of higher order in the (small) corrections $\Delta F_n^{(m)}$ and $\Delta G_n^{(m)}$. Substituting (25) into the recursion formulas and using the fact that $F_n^{(m)}$ and $G_n^{(m)}$ satisfy the recursion formulas, we obtain

$$\begin{aligned}
\Delta F_n^{(m)} &= \frac{1}{n^2-1}\left[-\mu^2 \Delta G_{n-2}^{(m)} + \gamma \sum_{j=1}^{n-2} \left(F_j^{(m)} \Delta G_{n-j-1}^{(m)} + G_{n-j-1}^{(m} \Delta F_j^{(m)}\right)\right], \\
\Delta G_n^{(m)} &= \frac{1}{n^2-1}\left[\mu^2 \Delta F_{n-2}^{(m)} - \gamma \sum_{j=1}^{n-2} F_j^{(m)} \Delta F_{n-j-1}^{(m)}\right], \qquad n = 3, 5, \cdots.
\end{aligned} \tag{26}$$

Now equations (26) are simply recursion formulas for the corrections $\Delta F_n^{(m)}$ and $\Delta G_n^{(m)}$. Hence, by $(n-1)/2$ applications of (26), $\Delta F_n^{(m)}$ and $\Delta G_n^{(m)}$ can be written as linear combinations of $\Delta F_1^{(m)}$ and $\Delta G_1^{(m)}$, thus

$$\begin{aligned} \Delta F_n^{(m)} &= a_n^{(m)} \Delta F_1^{(m)} + b_n^{(m)} \Delta G_1^{(m)}, \\ \Delta G_n^{(m)} &= c_n^{(m)} \Delta F_1^{(m)} + d_n^{(m)} \Delta G_1^{(m)}. \end{aligned} \tag{27}$$

where

$$\begin{aligned} a_1^{(m)} &= 1, \quad b_1^{(m)} = 0, \quad c_1^{(m)} = 0, \quad d_1^{(m)} = 1, \\ a_n^{(m)} &= \frac{1}{n^2-1}\left[-\mu^2 c_{n-2}^{(m)} + \gamma \sum_{j=1}^{n-2} (F_j^{(m)} c_{n-j-1}^{(m)} + G_{n-j-1}^{(m)} a_j^{(m)})\right], \\ b_n^{(m)} &= \frac{1}{n^2-1}\left[-\mu^2 d_{n-2}^{(m)} + \gamma \sum_{j=1}^{n-2} (F_j^{(m)} d_{n-j-1}^{(m)} + G_{n-j-1}^{(m)} b_j^{(m)})\right], \\ c_n^{(m)} &= \frac{1}{n^2-1}\left[\mu^2 a_{n-2}^{(m)} - \gamma \sum_{j=1}^{n-2} F_j^{(m)} a_{n-j-1}^{(m)}\right], \\ d_n^{(m)} &= \frac{1}{n^2-1}\left[\mu^2 b_{n-2}^{(m)} - \gamma \sum_{j=1}^{n-2} F_j^{(m)} b_{n-j-1}^{(m)}\right], \qquad n = 3, 5, \cdots. \end{aligned} \tag{28}$$

As a second requirement on $F_n^{(m+1)}$ and $G_n^{(m+1)}$, we require that

$$\sum_n F_n^{(m+1)} x^n \quad \text{and} \quad \sum_n G_n^{(m+1)} x^n$$

satisfy the boundary conditions (8). Using (27), the second requirement can be written

$$\Delta F_1^{(m)} \sum_n (n+\nu) a_n^{(m)} + \Delta G_1^{(m)} \sum_n (n+\nu) b_n^{(m)} = -\sum_n (n+\nu) F_n^{(m)}, \tag{29}$$

$$\Delta F_1^{(m)} \sum_n (n-\nu) c_n^{(m)} + \Delta G_1^{(m)} \sum_n (n-\nu) d_n^{(m)} = -\sum_n (n-\nu) G_n^{(m)}. \tag{30}$$

Notice that the non-homogeneous terms in equations (29) and (30) are the amounts by which the boundary conditions failed to be satisfied by the m^{th} approximation. In other words, the non-homogeneous terms are proportional to the boundary values of M_ξ and u implied by the m^{th} approximation. There remains only to solve equations (29) and (30) for $\Delta F_1^{(m)}$ and $\Delta G_1^{(m)}$ to obtain the $(m+1)^{\text{th}}$ approximation of F_1 and G_1 given by (25).

6. Extension to the case of a clamped edge. We now consider what changes in Sections 4 and 5 are necessary if the edge of the shell is clamped rather than simply supported. For a clamped edge, all preceding equations remain the same as for a simply supported edge except that the following equations numbered with a prime must be used in place of the corresponding unprimed equations.

$$\begin{aligned} F_1^{(0)} &= \frac{1-\nu}{\sqrt{2}\mu} \frac{ber_1\mu + bei_1\mu}{\mu(ber_1'\mu ber_1\mu + bei_1'\mu bei_1\mu) - \nu(ber_1^2\mu + bei_1^2\mu)} \\ G_1^{(0)} &= \frac{1-\nu}{\sqrt{2}\mu} \frac{ber_1\mu - bei_1\mu}{\mu(ber_1'\mu ber_1\mu + bei_1'\mu bei_1\mu) - \nu(ber_1^2\mu + bei_1^2\mu)} + \frac{2}{\mu^2} \end{aligned} \tag{16'}$$

$$F = -\tfrac{1}{4}(x - x^3) \qquad G = 0 \tag{17'}$$

$$F_1^{(0)} = -\tfrac{1}{4} \qquad G_1^{(0)} = 0 \tag{18'}$$

$$C_1 = 2(4 - \nu) \qquad C_5 = 576(1 - \nu)$$

$$C_2 = -3(13 - 3\nu) \qquad C_6 = -\frac{3 - \nu}{16(1 - \nu)} \tag{23'}$$

$$C_3 = 9(5 - \nu)$$

$$C_4 = 2304(1 - \nu) \qquad C_7 = \frac{3 - \nu}{48(1 - \nu)}$$

$$\Delta F_1^{(m)} \sum_n a_n^{(m)} + \Delta G_1^{(m)} \sum_n b_n^{(m)} = -\sum_n F_n^{(m)} \tag{29'}$$

7. A Numerical example. We shall give a numerical example of the power series method of solution for a simply supported edge for the case $\gamma = \mu = 1$, $\nu = \frac{1}{3}$. Using the averaging method of Section 5, we find for the present example initial estimates of F_1 and G_1 from Tables 2 and 3:

$$F_1^{(0)} = -.525, \qquad G_1^{(0)} = .255. \tag{31}$$

Using the recursion formulas (14), we now find $F_n^{(0)}$ and $G_n^{(0)}$ for $n = 3, 5, 7, 9$, as listed in Table 4, that is, we approximate the power series solution by a ninth degree polynomial.

The number of terms of the power series it is necessary to retain in the solution, for practical purposes, depends on the precision of the numerical results desired. In our calculations, the first term that is omitted is the eleventh degree term, and it is sufficiently small that it does not affect the fourth significant figure in F or G. If greater precision is desired, that is if more significant figures are to be retained in the final result, then of course a higher degree polynomial must be used to approximate the power series.

TABLE 4

Values of $F_n^{(0)}$ and $G_n^{(0)}$ for numerical example

	n				
	1	3	5	7	9
$F_n^{(0)}$	−.525	.201 391	.007 404	−.000 715	.000 026
$G_n^{(0)}$	.255	−.082 852	.012 797	−.000 187	−.000 023

TABLE 5

Values of $F_n^{(1)}$ and $G_n^{(1)}$ for numerical example

	n				
	1	3	5	7	9
$F_n^{(1)}$	−.532 055	.202 404	.007 471	−.000 729	.000 026
$G_n^{(1)}$	.248 532	−.084 200	.012 920	−.000 188	−.000 033

As a measure of how well the ninth degree polynomial, with leading coefficients given by (31), satisfies the boundary condition $M_\xi(1) = 0$, we compare the value of the bending couple M_ξ at the edge with its value at the center of the shell. We cannot compare the value of u at the edge with its value at the center because its value at the center is zero. Hence, as a measure of how well the boundary condition $u(1) = 0$ is satisfied, we shall compare the value of $u(x)/x$ at the edge with its value at the center. We obtain

$$\left|\frac{M_\xi(1)}{M_\xi(0)}\right| = .0083, \qquad \left|\frac{[u(x)/x]_{x=1}}{[u(x)/x]_{x=0}}\right| = .043. \tag{32}$$

We now apply the method of Section 5 to obtain corrections to the leading coefficients in order that the boundary conditions will be more nearly satisfied. Specifically, we evaluate $a_n^{(0)}$, $b_n^{(0)}$, $c_n^{(0)}$, $d_n^{(0)}$ from (28), making use of the values of $F_n^{(0)}$ and $G_n^{(0)}$ already evaluated in the initial approximation. Then, using these quantities, we obtain the coefficients in the simultaneous linear equations (29) and (30) whose solution gives $\Delta F_1^{(0)}$ and $\Delta G_1^{(0)}$. Using the results of these calculations in (25), we find new estimates of the leading coefficients:

$$F_1^{(1)} = -.532\,055, \qquad G_1^{(1)} = .248\,532. \tag{33}$$

As before, a new series solution, with leading coefficients (33), is again approximated by a ninth degree polynomial whose coefficients are listed in Table 5.

With this solution, the measures of how well the boundary conditions are satisfied become:

$$\left|\frac{M_\xi(1)}{M_\xi(0)}\right| = .00002, \qquad \left|\frac{[u(x)/x]_{x=1}}{[u(x)/x]_{x=0}}\right| = .0005. \tag{34}$$

Comparison of (32) and (34) shows that one refinement of the leading coefficients results in a great improvement in the satisfaction of the boundary conditions.

For the numerical example $\gamma = \mu = 1$, $\nu = \frac{1}{3}$, it may be of interest to compare the three methods of obtaining the initial estimate of the leading coefficients discussed in Section 4.

First, as the crudest initial estimate, one could use the solution of linear plate theory, for which $\gamma = \mu = 0$, and from (18) one has

$$F_1^{(0)} = -.625, \qquad G_1^{(0)} = 0. \tag{35}$$

Second, as a better initial estimate, one could use the solution of linear shell theory, for which $\gamma = 0$, and from Table 1, one obtains

$$F_1^{(0)} = -.568, \qquad G_1^{(0)} = .219. \tag{36}$$

Finally, the averaging process, taking account of non-linearity, provides the best initial estimate, and it was found from Tables 2 and 3 to be

$$F_1^{(0)} = -.525, \qquad G_1^{(0)} = .255. \tag{37}$$

If one takes as the correct values of F_1 and G_1 the estimate (33), one has

$$F_1 = -.532, \qquad G_1 = .248. \tag{38}$$

By comparing (38) with (35), (36) and (37), one can see the relative accuracy of each of the three methods of obtaining initial estimates of the leading coefficients, and in particular, one sees that the averaging method yields the best initial estimate. It may be remarked that for larger values of $|\gamma|$ and μ than used in the present numerical example, the advantage of using the third method over either of the other two may be even more pronounced than here. However, further numerical work must be carried out before an estimate is possible in regard to the complete domain of γ and μ-values for which the present method is practical.

8. Stability. An increase of center deflection of the shell corresponds to an increase of pressure, provided either that the pressure is directed outwardly ($\gamma > 0$), or if directed inwardly ($\gamma < 0$), it is sufficiently small. However, for sufficiently large inward pressure, the possibility exists that an increase of center deflection accompanies a decrease of pressure. The smallest inward pressure for which this can occur may be called the critical pressure, and the corresponding state of deformation may be said to be unstable.

Let $W = w(0)/h$ be the ratio of the center deflection to thickness of the shell. W depends on the load parameter γ and the geometrical parameter μ. In terms of W, the condition for this type of instability is

$$\frac{\partial W}{\partial \gamma} = \infty, \tag{39}$$

and the smallest value of γ that satisfies (39) is the critical value of γ, denoted by γ_{cr}.

From (15), W can be expressed in terms of the coefficients of the power series solution as follows:

$$W = \frac{\gamma}{m^2} \sum \frac{F_n}{n+1}. \tag{40}$$

From (40),

$$\frac{\partial W}{\partial \gamma} = \frac{W}{\gamma} + \frac{\gamma}{m^2} \sum \frac{1}{n+1} \frac{\partial F_n}{\partial \gamma}. \tag{41}$$

Since W is finite, the requirement $\partial W/\partial \gamma = \infty$ is equivalent to the requirement that

$$\sum \frac{1}{n+1} \frac{\partial F_n}{\partial \gamma} = \infty. \tag{42}$$

In order to obtain a first approximation to the critical pressure, we consider the approximation (19) for F: $F_k = Ax + Bx^3$. Both the boundary condition (8) for a simply supported edge and the boundary condition (9) for a clamped edge

require that B be proportional to A. Hence, using the above polynomial approximation for F, we write in place of (42)

$$\frac{\partial A}{\partial \gamma} = \infty, \tag{43}$$

which, according to (21), is equivalent to the following requirement:

$$3C_1\gamma^2A^2 + 2C_2\gamma\mu^2A + C_3\mu^4 + C_4 = 0. \tag{44}$$

We eliminate A between (21) and (44) and solve the resultant equation explicitly for an estimate of the critical value of γ:

$$\gamma_{CR} = -\frac{1}{27C_1^2C_5}\{C_2[(2C_2^2 - 9C_1C_3)\mu^4 - 9C_1C_4]\mu^2 - 2[(C_2^2 - 3C_1C_3)\mu^4 - 3C_1C_4]^{\frac{3}{2}}\}. \tag{45}$$

This quantity is real only if

$$\mu^4 \geqq \frac{3C_1C_4}{C_2^2 - 3C_1C_3}, \tag{46}$$

hence, according to the approximation used above, instability can occur only if the inequality (46) is satisfied. That is, instability cannot occur for a flat plate nor for a shell that is too shallow.

For $\nu = \frac{1}{3}$, (46) requires that $\mu \geqq 2.1$ for a simply supported edge, and $\mu \geqq 3.1$ for a clamped edge. The critical values of γ that correspond to the minimum values of μ for which instability can occur are, for a simply supported edge, $\gamma_{CR} = -8.5$, and for a clamped edge, $\gamma_{CR} = -72$. We note that the classical buckling pressure for a complete sphere, in our notation, is $\gamma_{CR} = -\mu^4$. The critical pressures above are both less than the classical buckling pressure.

We have obtained an estimate only of the critical value of γ and the smallest value of μ for which the type of instability considered here can exist. In their paper concerning a perturbation type solution of the present problem, Kaplan and Fung [2] obtain the buckling pressure for a shell with a clamped edge on the basis of a two term perturbation series in powers of deflection over thickness. In [2] it is stated that the minimum value of μ* for which γ_{CR} exists is that for which the critical deflection (the center deflection for $\gamma = \gamma_{CR}$) equals the initial height of the shell. According to Fig. 13, of the above reference, the minimum value of μ for which buckling can occur is $\mu = 2.1$.

The critical pressure of [2] is given in terms of the quantity

$$P = -\frac{\gamma}{3m^2}\left(\frac{\sin\alpha}{\alpha}\right)^4 \cong -\frac{\gamma}{3m^2}.$$

For $\mu = 3.1$, Fig. 13 of [2] indicates $P_{cr} = 4.6$. In order to compare this critical

* Actually, in [2], the parameter $\lambda = \frac{1}{\alpha}\sqrt{2(1 - \cos\alpha)}\mu$ is used rather than μ. However, for a shallow shell, $\alpha << \pi/2$, and $\lambda \simeq \mu$.

pressure with that given by our approximation for $\mu = 3.1$, we compute the value of γ that corresponds to the above value of P_{cr}: $-3m^2P_{cr} = -45$.

We conclude that our numerical results for a clamped edge do not agree very closely with those of [2]. This lack of agreement is not surprising since the critical pressure was derived on the basis of the polynomial approximation (19) which can be expected to be a reasonably good approximation only for small values of γ and μ. It remains to carry out further, more accurate calculations for the stability problem, on the basis of the power series procedure of the present paper.

REFERENCES

1. Federhofer, K., *Elastizitat der flachen Kugelschale bei gleichförmigem Oberflächendrucke*, Österr. Wochenschrift fur den öffentl. Baudienst vol. 25 and 26 (1916).
2. Kaplan, A. and Fung, Y. C., *A nonlinear theory of bending and buckling of thin elastic shallow spherical shells*, N.A.C.A., T.N. 3212 (1954).
3. Reissner, E., *On axisymmetrical deformations of thin shells of revolution*, Proceedings of Symposia in Applied Mathematics, American Mathematical Society, vol. 3 (1950) p. 27.
4. Way, S., *Bending of circular plates with large deflection*, Transactions of the American Society of Mechanical Engineers, vol. 56 (1934) p. 627.

International Business Machines Corporation
San Francisco, California

(Received September 23, 1955)

SOME DEFINITE INTEGRALS INVOLVING CONICAL FUNCTIONS

By Leopold B. Felsen

An analysis of the problem of diffraction of a scalar or vector (electromagnetic) plane wave by a semi-infinite cone in terms of eigenfunctions in the $r\phi$ domain in a spherical coordinate system leads to a series solution involving certain integrals of conical functions.[1] For small cone angles some pertinent integrals are:

$$I_m(\theta, \theta') = \int_0^\infty dx\, f(x) K_x^m(\cos\theta) K_x^m(\cos\theta'), \qquad f(x) = x \frac{\tanh \pi x}{\cosh \pi x}, \tag{1}$$

$$J_m(\theta, \theta') = \int_0^\infty dx\, f(x)\varphi(x) K_x^m(\cos\theta) K_x^m(\cos\theta'), \qquad \varphi(x) = (\tfrac{1}{4} + x^2), \tag{2}$$

$$M_1(\theta, \theta') = \int_0^\infty dx\, f(x) \frac{K_x^1(\cos\theta) K_x^1(\cos\theta')}{\varphi(x)}, \tag{3}$$

where $m = 0, 1$. $K_x^m(\cos\theta) = P_{-\frac{1}{2}+ix}^m(\cos\theta)$ is the conical function of order x, degree m and argument $\cos\theta$. In view of the asymptotic behavior of K_x^m,[2]

$$K_x^m(\cos\theta) \sim \frac{1}{(2\pi \sin\theta)^{\frac{1}{2}}} (ix)^{m-\frac{1}{2}} e^{x\theta + i(\pi/4 + m\pi/2)}, \qquad x \to \infty, \qquad 0 < \theta < \pi, \tag{4}$$

the above integrals exist and converge uniformly in θ and θ' if $\theta + \theta' < \pi - \delta$, $\delta > 0$. In all subsequent considerations these restrictions on θ and θ' are assumed.

The evaluation of the above and other similar integrals can be carried out by employing the differential equation for K_x^m,

$$\left[L(\theta) - \frac{m^2}{\sin^2\theta}\right] K_x^m(\cos\theta) = (\tfrac{1}{4} + x^2) K_x^m(\cos\theta), \tag{5}$$

$$L(\theta) = \frac{1}{\sin\theta} \frac{d}{d\theta} \sin\theta \frac{d}{d\theta}, \tag{5a}$$

and an extension of a solution obtained by Mehler:[3]

$$\frac{1}{w + z} = \pi \int_0^\infty dx\, f(x) K_x(w) K_x(z), \qquad 1 \leqq w < \infty, \qquad -1 < z < \infty. \tag{6}$$

If we let $w = \cos\theta'$, $z = \cos\theta$, the integral in (6) converges also if $w = \cos\theta' < 1$, $\theta + \theta' < \pi$. Since

$$K_x(w) = F\left(\tfrac{1}{2} - ix, \tfrac{1}{2} + ix; 1; \frac{1 - w}{2}\right) \tag{7}$$

can be continued analytically into the complex w-plane for $|1 - w| < 2$, and since the integral in (6) converges if $|1 - w| < 1 + \cos\theta$, it follows, upon also

[1] L. B. Felsen, "Plane Wave Scattering by Small Angle Cones," Transactions IRE-PGAP, October 1956.

[2] W. Magnus and F. Oberhettinger, "Special Functions of Mathematical Physics," Chelsea Pub. Co., New York, 1949, p. 71.

[3] F. G. Mehler, Math. Ann., Vol. 18, 1881, Sec. 4, Eq. (17).

continuing the left-hand side analytically into the complex w-plane that (6) is valid also for $-1 < w \leqq 1$, $-1 < z \leqq 1$, provided that $(w + z) > 0$. Thus,

$$I_0(\theta, \theta') = \frac{1}{\pi(\cos\theta + \cos\theta')}, \qquad \theta + \theta' < \pi. \tag{8}$$

Since $(d/d\theta)K_x(\cos\theta) = K_x^1(\cos\theta)$, one obtains directly:

$$I_1(\theta, \theta') = \frac{\partial^2}{\partial\theta\,\partial\theta'} I_0(\theta, \theta') = \frac{2}{\pi} \frac{\sin\theta \sin\theta'}{(\cos\theta + \cos\theta')^3}. \tag{9}$$

The interchange of differentiation and integration in (1) is permissible since all integrals converge uniformly in θ and θ' under the above restrictions.

In view of the differential equation (5) we have:

$$J_0(\theta, \theta') = L(\theta) I_0(\theta, \theta') = \frac{2}{\pi} \frac{1 + \cos\theta \cos\theta'}{(\cos\theta + \cos\theta)^3}. \tag{10}$$

Upon combining the results for I_0 and J_0 one obtains directly:

$$\int_0^\infty dx\, f(x) x^2 K_x(\cos\theta) K_x(\cos\theta') = J_0(\theta, \theta') - \tfrac{1}{4} I_0(\theta, \theta'). \tag{11}$$

By repeated differentiation as in (10) one obtains integrals containing integral powers of $\varphi(x)$. By expansion and subsequent regrouping, integrals as in (11), with x^2 replaced by x^{2n}, n integer, can be evaluated. Similarly,

$$\begin{aligned} J_1(\theta, \theta') &= \left(L(\theta) - \frac{1}{\sin^2\theta}\right) I_1(\theta, \theta') \\ &= \frac{4}{\pi} \frac{\sin\theta \sin\theta'}{(\cos\theta + \cos\theta')^4} (6 + 4\cos\theta\cos\theta' - \cos^2\theta - \cos^2\theta'). \end{aligned} \tag{12}$$

The extension to integrands involving x^{2n} is effected as above.

To evaluate $M_1(\theta, \theta')$, we employ:

$$L(\theta) M_0(\theta, \theta') = I_0(\theta, \theta'), \tag{13}$$

where M_0 is the integral in (3), with K_x^1 replaced by K_x. Upon integrating (13) over θ between the limits 0 and θ one obtains:

$$\sin\theta \frac{dM_0}{d\theta}\Big]_0^\theta = \int_0^\theta d\theta \sin\theta\, I_0 = -\frac{1}{\pi} \ln(\cos\theta + \cos\theta')\Big]_0^\theta. \tag{14}$$

Since $dM_0/d\theta$ is finite at $\theta = 0$ one obtains:

$$\frac{dM_0}{d\theta} = -\frac{1}{\pi \sin\theta} \ln\left(\frac{\cos\theta + \cos\theta'}{1 + \cos\theta'}\right). \tag{15}$$

Thus,

$$M_1(\theta, \theta') = \frac{\partial^2}{\partial\theta'\,\partial\theta} M_0(\theta, \theta') = \frac{1}{\pi} \frac{\tan(\theta/2) \tan(\theta'/2)}{\cos\theta + \cos\theta'}. \tag{16}$$

MICROWAVE RESEARCH INSTITUTE
POLYTECHNIC INSTITUTE OF BROOKLYN

(Received May 26, 1955)

COMPRESSIBLE VISCOUS FLOW PAST A WEDGE MOVING AT HYPERSONIC SPEEDS*

By L. J. Pan and Y. H. Kuo

1. Introduction. When a thin sharp-nosed body flies at hypersonic speeds, the flow around the body is characterized by its high temperature and low density near the surface, due largely to the loss of kinetic energy of the fluid by the action of viscosity. The consequence of expansion of the gas in the boundary layer is an enormous increase of the viscous deflection of the outer flow, which in turn causes a strong shock at the leading edge. Previous investigations (Refs. 1, 2, 3) have laid special emphasis on the increase of the effective thickness of the body due to viscous deflection of the stream-lines but have neglected the entropy or vorticity generated by the shock.

In view of the fact that at hypersonic Mach numbers the change across the shock, even in the case of thin bodies, can be large, an increase of entropy could be accompanied by an appreciable increase of temperature of the flow. This heating may lead to further expansion of the gas, resulting in additional changes of pressure and skin friction. If, however, these effects were taken into account, the flow behind the shock would involve the interaction of both families of Mach waves and, therefore, would present formidable difficulties for theoretical solution. On the other hand, if the basis for the simple-wave theory can be established, the solution of the problem can be greatly simplified as has been demonstrated in the case of the flat plate (Ref. 4). It is, for this reason, important to clarify to what extent entropy will be of significance.

With a motivation to answer this question, it is proposed to consider a special case of an infinite wedge with large wedge angle, for which a strong shock exists even at infinity. Inasmuch as, at high Mach numbers the characteristics of the flow over thin sharp-nosed bodies depend very little on the shape of the body, the results obtained for one case may well cast light on the other. Merely for the purpose of estimating these effects, the investigations can be restricted to the weak interaction zone for which an asymptotic solution can be easily found. It is shown that at hypersonic Mach numbers the contributions due to entropy rise are small even for large wedge angles. It can therefore be concluded that, as in the case of inviscid rotational flows (Ref. 5), the waves reflected by the shock play only a minor role in influencing the flow and can be neglected at hypersonic Mach numbers.

2. Equations of motion. In the Cartesian system $(\bar{x}, \bar{y})$, let the surface of the wedge be defined by $0 \leq \bar{x}$, and $\bar{y} = 0$. The flow field in question is supersonic and is bounded by the leading-edge shock $0S$ (Fig. 1) and the surface

* This work, being part of a thesis by L. J. Pan, was carried out at the Graduate School of Aeronautical Engineering under the sponsorship of the Mechanics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

of the wedge. Let $\bar{u}$, $\bar{v}$ be the velocities in the $\bar{x}$, $\bar{y}$ directions respectively and let $\bar{p}$, $\bar{\rho}$, $\bar{T}$ and $\bar{\mu}$ be, respectively, the pressure, density, temperature and viscosity coefficient. If the velocity U_∞, the thermal variables p_∞, ρ_∞, T_∞ and the viscosity μ_∞ far downstream are taken as references, the variables can be normalized as follows:

$$\left.\begin{aligned} x &= \bar{x}/L, & Y &= (Re/C)^{\frac{1}{2}}\bar{y}/L = (Re/C)^{\frac{1}{2}}y \\ u &= \bar{u}/U_\infty, & V &= (Re/C)^{\frac{1}{2}}\bar{v}/U_\infty = (Re/C)^{\frac{1}{2}}v, \quad p = \bar{p}/p_\infty \\ \rho &= \bar{\rho}/\rho_\infty, & T &= \bar{T}/T_\infty, \qquad \mu = \bar{\mu}/\mu_\infty \end{aligned}\right\} \tag{2.1}$$

where $Re = \rho_\infty U_\infty L/\mu_\infty$. Here L is an arbitrary length and C is a constant to be defined later. It is remarked, however, that the scale L is introduced only as a convenience in formulation and will drop out in the final solutions.

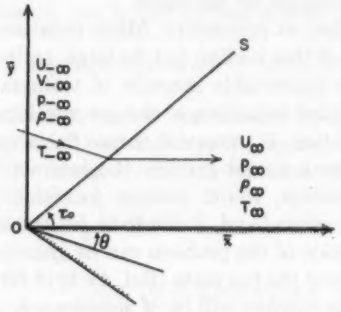


FIG. 1. Diagram showing the unperturbed supersonic flow over infinite wedge

In terms of these variables, the Navier-Stokes equations for a compressible viscous fluid can be expressed as:

$$\left.\begin{aligned} \rho\frac{Du}{Dt} &= -\left[\frac{p}{\gamma M^2} + \frac{2}{3Re}\mu\Theta\right]_x + \frac{1}{Re}\{2(\mu u_x)_x + [\mu(u_y + v_x)]_y\} \\ \rho\frac{Dv}{Dt} &= -\left[\frac{p}{\gamma M^2} + \frac{2}{3Re}\mu\Theta\right]_y + \frac{1}{Re}\{[\mu(u_y + v_x)]_x + 2(\mu v_y)_y\} \\ (\rho u)_x &+ (\rho v)_y = 0 \end{aligned}\right\} \tag{2.2}$$

with

$$\frac{D}{Dt} = u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} \qquad \textit{and} \qquad \Theta = u_x + v_y$$

where γ is the ratio of the specific heats, $M = U_\infty/a_\infty$, a_∞ being the speed of sound at positive infinity, and the subscripts denote partial differentiation with respect to the variables indicated.

In the case of gases, the above equations involving both ρ and T by the dependency on temperature of the viscosity, cannot form a complete system with-

out consideration of the energy balance and the equation of state. For Prandtl number unity, these are, respectively,

$$\rho \frac{DT}{Dt} = \frac{\gamma - 1}{\gamma} \frac{Dp}{Dt} + \frac{(\gamma - 1)M^2}{Re} [-\tfrac{2}{3}\mu\Theta^2 + 2\mu(u_x^2 + v_y^2) + \mu(u_y + v_x)^2] + \frac{1}{Re} [(\mu T_x)_x + (\mu T_y)_y] \tag{2.3}$$

and, for perfect gases,

$$p = \rho T \tag{2.4}$$

In the following, it is proposed to attack this problem approximately by two different methods according to the distance from the surface. At large y, the deviation of the flow properties from those of the uniform state at infinity is small and the solution can be obtained by perturbation, taking the uniform flow as the zero approximation. This region will be subsequently called *external field*. Near the surface, where the viscous effect is important, boundary-layer approximations will be employed. This region will be called the *boundary layer*.

3. First-order solution of the external field. As the non-uniformity of the flow behind the shock is caused primarily by the viscous displacement of the streamlines, the velocity in the external field must join, to the first approximation, with the velocity produced by the boundary layer along a flat plate, which is

$$\sqrt{\frac{C}{Re}} \frac{0.860 + 0.596(\gamma - 1)M^2}{\sqrt{x}}$$

Thus, at $y = 0$, the external field must satisfy:

$$v = \epsilon v_1/\sqrt{x} \tag{3.1}$$

where

$$\epsilon = \sqrt{C/Re} \quad \text{and} \quad v_1 = 0.860 + 0.596(\gamma - 1)M^2$$

The parameter ϵ so defined can serve as a measure of the first-order viscous effect.

With this definition of ϵ, the following expansions can be attempted: for $\epsilon \ll 1$

$$\left.\begin{aligned} u &= 1 + \epsilon u^{(1)} + \epsilon^{\frac{1}{2}} u^{(\frac{1}{2})} + \cdots \\ v &= \quad \epsilon v^{(1)} + \epsilon^{\frac{1}{2}} v^{(\frac{1}{2})} + \cdots \\ p &= 1 + \epsilon p^{(1)} + \epsilon^{\frac{1}{2}} p^{(\frac{1}{2})} + \cdots \\ \rho &= 1 + \epsilon \rho^{(1)} + \epsilon^{\frac{1}{2}} \rho^{(\frac{1}{2})} + \cdots \\ T &= 1 + \epsilon T^{(1)} + \epsilon^{\frac{1}{2}} T^{(\frac{1}{2})} + \cdots \end{aligned}\right\} \tag{3.2}$$

By substituting the expansions (3.2) in Eqs. (2.2), (2.3) and (2.4), it can be readily seen that the first-order solutions satisfy a differential system which

is identically the same as the first-order inviscid differential system. Solutions of such a system have been already obtained by Chu (Ref. 6); these are as follows:

$$\left.\begin{aligned} v^{(1)} &= v_1[F_1(x-\beta y)+G_1(x+\beta y)] \\ p^{(1)} &= (\gamma M^2 v_1/\beta)[F_1(x-\beta y)-G_1(x+\beta y)] \\ u^{(1)} &= -(v_1/\beta)[F_1(x-\beta y)-G_1(x+\beta y)]-S_1(y) \\ \rho^{(1)} &= (M^2 v_1/\beta)[F_1(x-\beta y)-G_1(x+\beta y)]-(\gamma-1)M^2 S_1(y) \\ T^{(1)} &= [(\gamma-1)M^2 v_1/\beta]\,[F_1(x-\beta y)-G_1(x+\beta y)]+(\gamma-1)M^2 S_1(y) \end{aligned}\right\} \tag{3.3}$$

where $\beta = \sqrt{M^2-1}$; F_1, G_1 and S_1 are arbitrary functions.

This shows that the first-order solutions consist of both the outgoing and the reflected families of plane waves and, in addition, a "vorticity wake." From the definition of entropy, viz. $\ln p\rho^{-\gamma}$, $S_1(y)$ is simply proportional to the first-order entropy which is, therefore, constant on $y =$ constant lines. It is interesting to note that the effect of the entropy generated by the shock, to the first order, is to decrease the velocity and density but to increase the temperature, as compared with an irrotational flow.

The flow field described above can now be determined jointly with the shape of the shock by satisfying the Rankine-Hugoniot shock relations and the condition (3.1). As the thickness of the shock is much smaller than the thickness of the boundary layer, the shock, to the first approximation, can be described by a curve:

$$dy/dx = \tan\tau \tag{3.4}$$

where τ is the local shock angle. The expansion of the shock conditions according to (3.2), together with

$$\tau = \tau_0 + \epsilon\tau_1 + \cdots \tag{3.5}$$

where τ_0 is the shock angle at infinity and is determined by the half wedge angle θ and the Mach number $M_{-\infty}$ in front of the shock, leads to a set of first-order relations (Ref. 6) on the shock $y = x\tan\tau_0$. From these it follows, by elimination that

$$p^{(1)} = \gamma M^2 v^{(1)} \tan\lambda \tag{3.6}$$

$$\tau^{(1)} = \tfrac{1}{2}(\gamma+1)\csc 2\tau_0\, p^{(1)}/\gamma M^2 \tag{3.7}$$

$$S_1 = \tfrac{1}{2}\frac{(1-M^2\sin^2\tau_0)^2}{\sin^2\tau_0[1+\frac{1}{2}(\gamma-1)M^2\sin^2\tau_0]}\frac{p^{(1)}}{\gamma M^4} \tag{3.8}$$

where

$$\tan\lambda = \frac{M^2[1+\frac{1}{2}(\gamma-1)M^2\sin^2\tau_0]\sin 2\tau_0}{\beta^2[1+\frac{1}{2}(\gamma-1)M^2\sin^2\tau_0]+M^2[\gamma M^2\sin^2\tau_0-\frac{1}{2}(\gamma-1)]\cos^2\tau_0} \tag{3.9}$$

These relations will be valid as long as the flow is supersonic and the shock belongs to the weak family.

To satisfy the conditions (3.1) and (3.6), the functions F_1 and G_1 may have the forms:

$$F_1 = \frac{1 - A_1}{\sqrt{x - \beta y}}, \qquad G_1 = \frac{A_1}{\sqrt{x + \beta y}} \tag{3.10}$$

The constant A_1, by (3.6), is determined to be

$$A_1 = \frac{(1 - \beta \tan \lambda)}{(1 - \beta \tan \lambda) + \sqrt{(1 - \beta \tan \tau_0)/(1 + \beta \tan \tau_0)}\,(1 + \beta \tan \lambda)} \tag{3.11}$$

As soon as the pressure $p^{(1)}$ is known, it follows from (3.7) and (3.8) that

$$\tau^{(1)} = \frac{\gamma + 1}{2} \frac{v_1}{\beta \sin 2\tau_0} \left[\frac{1 - A_1}{\sqrt{1 - \beta \tan \tau_0}} - \frac{A_1}{\sqrt{1 + \beta \tan \tau_0}} \right] \frac{1}{\sqrt{x}} \tag{3.12}$$

and

$$S_1 = \frac{B_1}{\sqrt{y}} \tag{3.13}$$

where

$$B_1 = \tfrac{1}{2} \frac{(1 - M^2 \sin^2 \tau_0)^2}{\beta M^2 \sin^2 \tau_0 [1 + \tfrac{1}{2}(\gamma - 1) M^2 \sin^2 \tau_0]} \cdot \left[\frac{1 - A_1}{\sqrt{\cot \tau_0 - \beta}} - \frac{A_1}{\sqrt{\cot \tau_0 + \beta}} \right] \tag{3.14}$$

On account of (3.12), to the first order, the shock curve is:

$$y = x \tan \tau_0 + \frac{(\gamma + 1)\epsilon v_1}{\beta^2 \cos^2 \tau_0 \sin 2\tau_0} \cdot \left[\frac{1 - A_1}{\sqrt{1 - \beta \tan \tau_0}} - \frac{A_1}{\sqrt{1 + \beta \tan \tau_0}} \right] \sqrt{x} + \cdots \tag{3.15}$$

It is observed that the first-order solutions given above are singular at $y = 0$ due to the presence in the flow of non-uniform entropy. This implies that in the neighborhood of the edge of the boundary layer the velocity gradient will not be small, as is the case of an irrotational external flow, and consequently, a "sharp boundary" between the boundary layer and the external field ceases to exist. As a result of this gradual transition from the viscous boundary layer to the inviscid external field, the familiar concept of matching the boundary-layer solution to that of the external field must be radically modified. This new problem will be solved in Section 6.

4. Singular solution of the order $\epsilon^{\frac{1}{2}}$. In view of the singular term S_1, it is inevitable that singular terms of the type $y^{-l}x^m$ will appear in the high-order solutions at small y, where the exponents l and m are positive and l increases with the order of approximation. For instance, by substituting

$$u = 1 - \epsilon S_1 - \epsilon^2 S_2 - \cdots - \epsilon^n S_n - \cdots \tag{4.1}$$
$$v = \epsilon v_1/\sqrt{x}$$

in the first of Eqs. (2.2), leaving out the regular terms, one finds that the highest singular term in S_2 is $B_1 v_1 y^{-\frac{1}{2}} x^{\frac{1}{2}}$; and the highest singular term in S_3 is $(\frac{3}{4})B_1(1 + 2v_1^2)y^{-\frac{1}{2}}x$, etc. To continue the process, it can be shown that for the n^{th} highest singular term S_{1n} in S_n there is an equation:

$$S_{1(n+2)x} + (v_1/\sqrt{x})S_{1(n+1)y} = S_{1nyy}\,, \quad n = 1, 2, 3, \cdots \tag{4.2}$$

From the first three solutions, it can be generalized that Eq. (4.2) admits a solution:

$$S_{1n} = B_1 C_n y^{-n+\frac{1}{2}} x^{(n-1)/2}, \qquad S_{11} = S_1 \tag{4.3}$$

A substitution of (4.3) in Eq. (4.2) yields the recurrence formula: for $n = 1, 2, 3, \cdots$

$$C_{n+2} = [2/(n+1)]\,[v_1(n + \tfrac{1}{2})C_{n+1} + (n^2 - \tfrac{1}{4})C_n] \tag{4.4}$$

with

$$C_1 = 1, \qquad C_2 = v_1$$

This particular sequence on summation gives a solution in the form of (4.1):

$$u = 1 - B_1 \sum_1^\infty \epsilon^n C_n y^{-n+\frac{1}{2}} x^{(n-1)/2} + \cdots$$

This series has the property that when $y = 0(\epsilon)$, all terms become the same order, $\epsilon^{\frac{1}{2}}$, and the series converges if $\epsilon\sqrt{x} < y$. For this reason, when $y = 0(\epsilon)$, the complete solution for u is:

$$u = 1 - \sqrt{\epsilon}u^{(\frac{1}{2})} + \epsilon\tilde{u}^{(1)} + \cdots$$

where

$$u^{(\frac{1}{2})} = B_1 \sum_1^\infty \epsilon^{n-\frac{1}{2}} C_n y^{-n+\frac{1}{2}} x^{(n-1)/2} \tag{4.5}$$

and $\tilde{u}^{(1)}$, being the regular part of $u^{(\frac{1}{2})}$ defined in (3.3), is a linear combination of F_1 and G_1 .

As one proceeds to still higher approximations, singular terms like (4.3), which are of the order ϵ, $\epsilon^{\frac{1}{2}}$, etc., have to be considered. For the present study, which stops at the first order, this problem will not be pursued further.

5. Boundary-layer equations. Consider now the solution in the inner, viscous layer. In the case of large Reynolds number, the viscous layer will be thin and the boundary-layer theory a legitimate approximation. Accordingly, the equations (2.2) can be satisfied to:

$$\left.\begin{aligned} \rho[uu_x + Vu_Y] &= -(1/\gamma M^2)p_x + 1/C[\mu u_Y]_Y \\ 0 &= (1/\gamma M^2)p_Y \\ (\rho u)_x + (\rho V)_Y &= 0 \end{aligned}\right\} \tag{5.1}$$

by neglecting terms of the order ϵ^2. If the surface is thermally insulated, the energy equation admits a solution:

$$T = 1 + \tfrac{1}{2}(\gamma - 1)M^2(1 - u^2) \tag{5.2}$$

Furthermore, it is assumed that the viscosity-temperature relation is linear. Following Chapman and Rubesin (Ref. 7),

$$\mu = CT \tag{5.3}$$

Here the constant C is chosen such that μ at the surface agrees with the power law T^ω, ω being a constant for individual gases. This determines

$$C = T_w^{\omega-1} = [1 + \tfrac{1}{2}(\gamma - 1)M^2]^{\omega-1} \tag{5.4}$$

where T_w is the temperature of the surface.

On account of (5.3), Eqs. (5.1) can be further reduced by introducing Howarth's transformation (Ref. 8)

$$\Psi = \sqrt{p}\varphi(\xi, \eta), \qquad x = \xi, \qquad Y = \frac{1}{\sqrt{p}}\int_0^\eta T\, dz \tag{5.5}$$

where Ψ is the non-dimensional stream function defined by:

$$\rho u = \Psi_Y, \qquad \rho V = -\Psi_x \tag{5.6}$$

According to the transformation (5.5), the velocities u and V can be expressed in terms of φ, T, and p:

$$\begin{aligned} u &= \varphi_\eta \\ V &= -\frac{1}{\sqrt{p}}\left[T\varphi_\xi - \varphi_\eta\int_0^\eta T_\xi\, dz\right] + \frac{p'}{2p}\left(T\varphi + \varphi_\eta\int_0^\eta T\, dz\right) \end{aligned} \tag{5.7}$$

A substitution of u and V in (5.1), by making use of the equation of state, leads to an equation:

$$\varphi_\eta\varphi_{\xi\eta} - \varphi_\xi\varphi_{\eta\eta} = \frac{p'}{p}\left[\tfrac{1}{2}\varphi\varphi_{\eta\eta} - \frac{\gamma - 1}{2\gamma}(1 - \varphi_\eta^2) - \frac{1}{\gamma M^2}\right] + \varphi_{\eta\eta\eta} \tag{5.8}$$

where the pressure p is a function of ξ alone. The solution of (5.8) must satisfy the conditions:

$$\begin{aligned} \varphi = \varphi_\eta &= 0, \qquad \eta = 0 \\ \varphi_\eta &= U(\xi, \eta), \qquad \eta \to \infty \end{aligned} \tag{5.9}$$

where $U(\xi, \eta)$ will be specified later.

6. Solutions for the viscous layer. From (4.5), the velocity u near the edge of the boundary layer is, instead of (3.2),

$$u = 1 + \sqrt{\epsilon}\, u^{(\frac{1}{2})} + \epsilon\bar{u}^{(1)} + \cdots$$

for $y = 0(\epsilon)$. Accordingly, the solutions in the viscous layer take the form:

$$\begin{aligned} \varphi &= \varphi^{(0)} + \sqrt{\epsilon}\varphi^{(\frac{1}{2})} + \epsilon\varphi^{(1)} + \cdots \\ p &= 1 + \epsilon p^{(1)} + \epsilon^{\frac{1}{2}} p^{(\frac{1}{2})} + \cdots \end{aligned} \tag{6.1}$$

These expansions lead, from (5.7) and (5.2), to the expressions

$$\left.\begin{aligned} u &= \varphi_\eta^{(0)} + \sqrt{\epsilon}\varphi_\eta^{(\frac{1}{2})} + \cdots \\ V &= v^{(1)} + \sqrt{\epsilon}v^{(\frac{1}{2})} + \cdots \\ T &= T^{(0)} + \sqrt{\epsilon}T^{(\frac{1}{2})} + \cdots \end{aligned}\right\} \tag{6.2}$$

where $v^{(1)}$, $v^{(\frac{1}{2})}$, $T^{(0)}$ and $T^{(\frac{1}{2})}$ are defined by:

$$\left.\begin{aligned} v^{(1)} &= -T^{(0)}\varphi_\xi^{(0)} + \varphi_\eta^{(0)} \int_0^\eta T_\xi^{(0)}\, dz \\ v^{(\frac{1}{2})} &= -T^{(0)}\varphi_\xi^{(\frac{1}{2})} - T^{(\frac{1}{2})}\varphi_\xi^{(0)} + \varphi_\eta^{(0)} \int_0^\eta T_\xi^{(\frac{1}{2})}\, dz + \varphi_\eta^{(\frac{1}{2})} \int_0^\eta T_\xi^{(0)}\, dz \\ T^{(0)} &= 1 + \frac{\gamma - 1}{2} M^2(1 - \varphi_\eta^{2(0)}), \qquad T^{(\frac{1}{2})} = -(\gamma - 1)M^2\varphi_\eta^{(0)}\varphi_\eta^{(\frac{1}{2})} \end{aligned}\right\} \tag{6.3}$$

By expanding (5.8) and equating to zero the coefficients of ϵ^0 and $\epsilon^{\frac{1}{2}}$, there result

$$\varphi_\eta^{(0)}\varphi_{\xi\eta}^{(0)} - \varphi_\xi^{(0)}\varphi_{\eta\eta}^{(0)} - \varphi_{\eta\eta\eta}^{(0)} = 0 \tag{6.4}$$

and

$$\varphi_\eta^{(0)}\varphi_{\xi\eta}^{(\frac{1}{2})} + \varphi_{\xi\eta}^{(0)}\varphi_\eta^{(\frac{1}{2})} - \varphi_\xi^{(0)}\varphi_{\eta\eta}^{(\frac{1}{2})} - \varphi_{\eta\eta}^{(0)}\varphi_\xi^{(\frac{1}{2})} - \varphi_{\eta\eta\eta}^{(\frac{1}{2})} = 0 \tag{6.5}$$

The boundary conditions for $\varphi^{(0)}$ and $\varphi^{(\frac{1}{2})}$ are:

$$\left.\begin{aligned} \varphi^{(0)} = \varphi_\eta^{(0)} = \varphi^{(\frac{1}{2})} = \varphi_\eta^{(\frac{1}{2})} = 0, \qquad \eta = 0 \\ \varphi_\eta^{(0)} = 1, \qquad \varphi_\eta^{(\frac{1}{2})} = u^{(\frac{1}{2})}(\xi, \eta), \qquad \eta \to \infty \end{aligned}\right\} \tag{6.6}$$

where $u^{(\frac{1}{2})}(\xi, \eta)$ is required to equal the series defined by (4.5).

Thus, the zero-order solution, as expected, is simply the boundary-layer solution for a flat plate under uniform pressure. The solution of Eq. (6.4) according to Blasius is

$$\varphi^{(0)} = \sqrt{\xi} f_0(\zeta) \tag{6.7}$$

with $\zeta = \eta/2\sqrt{\xi}$. Eq. (6.4) then reduces to

$$f_0''' + f_0 f_0'' = 0 \tag{6.8}$$

with the boundary conditions:

$$\left.\begin{aligned} f_0 = f_0' = 0, \qquad \zeta = 0 \\ f_0' = 2, \qquad \zeta = \infty \end{aligned}\right\} \tag{6.9}$$

The solution is well-known. For ζ near the origin, the solution can be expanded in a power series

$$f_0 = \frac{\alpha_0 \zeta^2}{2!} - \frac{\alpha_0^2 \zeta^5}{5!} + \frac{11\alpha_0^3 \zeta^8}{8!} - \frac{375\alpha_0^4 \zeta^{11}}{11!} + \cdots \tag{6.10}$$

where $\alpha_0 = 1.328$. On the other hand, when ζ is large, the solution is asymptotically

$$f_0 \sim 2\zeta - \beta_0 + 0(e^{-\zeta^2}) \tag{6.11}$$

where $\beta_0 = 1.720$.

Given f_0 above, both $u^{(0)}$ and $v^{(1)}$ in (6.2) can be expressed in terms of f_0 and its derivatives. From the asymptotic behavior of f_0, it can easily be shown that $v^{(1)}(\xi, \eta)$ approaches the value (3.1) as ζ becomes infinite.

With f_0 so calculated, it follows that the solution $\varphi^{(\frac{1}{2})}$ must be

$$\varphi^{(\frac{1}{2})} = -B_1 \xi^{\frac{1}{2}} f_{\frac{1}{2}}(\zeta) \tag{6.12}$$

where the constant B_1 is defined in (3.14). Eq. (6.5) is, consequently, reduced to

$$2f_{\frac{1}{2}}''' + 2f_0 f_{\frac{1}{2}}'' + f_0' f_{\frac{1}{2}}' + f_0'' f_{\frac{1}{2}} = 0 \tag{6.13}$$

with the boundary conditions

$$\left.\begin{aligned} f_{\frac{1}{2}} &= f_{\frac{1}{2}}' = 0, && \zeta = 0 \\ f_{\frac{1}{2}}' &= 2\textstyle\sum_1^\infty C_n(2\zeta - \beta_0 + 2v_1)^{-n+\frac{1}{2}}, && \zeta \to \infty \end{aligned}\right\} \tag{6.14}$$

where C_n is defined by (4.4).

It should be pointed out that, as already mentioned in Section 3, owing to the strong entropy gradient near the surface of the wedge, the two velocity fields, if they are truly matched, should be matched for all their derivatives in a region of large ζ; $\zeta = 5$, say. This is actually what is prescribed by the second of (6.14). Finally, the problem will be completed after it is verified that the required asymptotic behavior is exactly the value of the external velocity given by (4.5).

First, by satisfying the conditions at $\zeta = 0$ the power-series development yields

$$f_{\frac{1}{2}} = \alpha_{\frac{1}{2}}\left[\frac{\zeta^2}{2!} - 3\alpha_0 \frac{\zeta^5}{5!} + 54\alpha_0^2 \frac{\zeta^8}{8!} - 2586\alpha_0^3 \frac{\zeta^{11}}{11!} + \cdots\right] \tag{6.15}$$

where $\alpha_{\frac{1}{2}}$ determined by numerical integration, is 1.83150. Secondly, when ζ is large, by approximating f_0 according to (6.11), Eq. (6.13) is simplified to

$$f_{\frac{1}{2}}''' + (2\zeta - \beta_0) f_{\frac{1}{2}}'' + f_{\frac{1}{2}}' = 0 \tag{6.16}$$

Upon introduction of a new variable, $\sigma = 2\zeta - \beta_0$, the resulting equation can be integrated in terms of Bessel functions of imaginary argument. Namely,

$$\frac{df_{\frac{1}{2}}}{d\sigma} = \sqrt{\sigma} e^{-\sigma^2/2}\left[D_1 I_{\frac{1}{4}}\left(\frac{\sigma^2}{2}\right) + D_2 K_{\frac{1}{4}}\left(\frac{\sigma^2}{2}\right)\right] \tag{6.17}$$

Now as $\sigma \to \infty$, asymptotically

$$I_{\frac{1}{4}}\left(\frac{\sigma^2}{2}\right) \sim \frac{e^{\sigma^2/2}}{\sqrt{\pi}} \sum_0^\infty \frac{(-1)^m \Gamma(\frac{3}{4} + m)}{m!\Gamma(\frac{3}{4} - m)} \sigma^{-2m-1} + 0(e^{-\sigma^2/2})$$

$$K_{\frac{1}{4}}\left(\frac{\sigma^2}{2}\right) \sim 0(e^{-\sigma^2/2})$$

It follows that as $\zeta \to \infty$ asymptotically

$$\frac{df_{\frac{1}{4}}}{d\zeta} \sim \frac{2D_1}{\sqrt{\pi}} \sum_0^\infty \frac{(-1)^m \Gamma(\frac{3}{4}+m)}{m!\Gamma(\frac{3}{4}-m)} \sigma^{-2m-1} + 0(e^{-\zeta^2/2}) \tag{6.18}$$

From the transformation (5.5), when ζ is large the relation between x, y and ζ, to the first approximation, is

$$\frac{y}{\epsilon\sqrt{x}} = 2\zeta - \beta_0 + 2v_1 \tag{6.19}$$

By writing $\sigma = (2\zeta - \beta_0 + 2v_1) - 2v_1$ and replacing in (6.18)

$$\left(1 - \frac{2v_1}{\sigma + 2v_1}\right)^{-2m-1} \quad \text{by} \quad \sum_0^\infty \frac{\Gamma(2m+1+l)}{l!\Gamma(2m+1)} (\sigma + 2v_1)^{-l}(2v_1)^l$$

there will result a double series which, in the domain of validity, can be converted into a single series by Cauchy's method of multiplication:

$$f'_{\frac{1}{4}} = \frac{2D_1}{\sqrt{\pi}} \sum_1^\infty C_n(2\zeta - \beta_0 + 2v_1)^{-n+\frac{1}{2}} + 0(e^{-\zeta^2}) \tag{6.20}$$

The coefficients C_n satisfy exactly the same recurrence formula (4.4) if $C_1 = 1$. Therefore, if $D_1 = \sqrt{\pi}$, the solution $f'_{\frac{1}{4}}$ will satisfy the imposed condition (6.14) at infinity, which by (6.19) is just $\xi^{\frac{1}{4}}u^{(1)}$ defined by (4.5).

It is interesting to observe that, although the boundary-layer solution does approach the prescribed inviscid field exponentially, the displacement boundary-layer thickness, nevertheless, cannot be defined due to the singular behavior of the inviscid solution at $y = 0$. Under such circumstances, it seems that the concept of the displacement thickness has to be dropped in favor of the well-defined quantity v. It must be pointed out also that in spite of the fact that the boundary-layer thickness, according to (5.5), is increased due to the presence of the entropy field, the displacement boundary-layer thickness, if defined, will be decreased, since the flow deflection is decreased, as will be shown later.

7. $v^{(1)}$ and $p^{(1)}$ of the external field. In the preceding sections, it was shown that the first-order entropy field produced by a curved shock will induce, inside the boundary layer, velocity (u), density and temperature increments of the order $\epsilon^{\frac{1}{2}}$ but a velocity (v) increment of the order $\epsilon^{\frac{1}{4}}$. This modification of the flow deflection will in turn generate new trains of waves in the external field. To calculate these quantities, the process of Section 3 can be repeated. According to (6.3), the velocity $v^{(1)}$ expressed in terms of f_0, $f_{\frac{1}{4}}$ is

$$v^{(1)} = \frac{B_1\xi^{-\frac{1}{2}}}{8}\left\{2(f_{\frac{1}{4}} - 2\zeta f'_{\frac{1}{4}}) + \frac{(\gamma-1)M^2}{2}\left[2\left(1 - \frac{f_0'^2}{4}\right)(f_{\frac{1}{4}} - 2\zeta f'_{\frac{1}{4}}) + f'_0(f''_{\frac{1}{4}} - \zeta f'_0 f'_{\frac{1}{4}} + \tfrac{1}{2}f'_0 f_{\frac{1}{4}} - \alpha_{\frac{1}{4}}) + f'_{\frac{1}{4}}(f''_0 - \alpha_0)\right]\right\} \tag{7.1}$$

when ζ is large, the asymptotic expansions of f_0 and $f_{\frac{1}{2}}$ lead to:

$$v^{(\frac{1}{2})} \sim -\frac{B_1 \xi^{-1}}{8}\left[(\gamma - 1)\, M^2(\alpha_{\frac{1}{2}} + \beta_{\frac{1}{2}}) + 2\beta_{\frac{1}{2}} + 0\left(\frac{1}{\zeta}\right)\right] \tag{7.2}$$

where $\beta_{\frac{1}{2}} = 0.1977$, being the constant defined as $f_{\frac{1}{2}} \sim -\beta_{\frac{1}{2}} + 0(\sqrt{\zeta})$ as ζ becomes large.

This negative value of $v^{(\frac{1}{2})}$ at infinity shows that the entropy field, contrary to the viscous effect, turns the flow toward the wall and, therefore, reduces the viscous displacement of the stream lines. There is an attendant reduction of pressure.

Following Section 3, if the asymptotic value in (7.2) is considered as the boundary value of the external field, the solutions must satisfy:

$$v^{(\frac{1}{2})} = -v_{\frac{1}{2}}/x^{\frac{1}{2}}, \qquad y = 0 \tag{7.3}$$

where

$$v_{\frac{1}{2}} = \tfrac{1}{8}B_1[2\beta_{\frac{1}{2}} + (\gamma - 1)M^2(\alpha_{\frac{1}{2}} + \beta_{\frac{1}{2}})]$$

Inasmuch as the shock relations to this order remain the same as those from (3.6) to (3.8), it follows therefore that $v^{(\frac{1}{2})}$ and $p^{(\frac{1}{2})}$ are,

$$\left.\begin{aligned} v^{(\frac{1}{2})} &= -v_{\frac{1}{2}}\left[\frac{1 - A_{\frac{1}{2}}}{(x - \beta y)^{\frac{1}{2}}} + \frac{A_{\frac{1}{2}}}{(x + \beta y)^{\frac{1}{2}}}\right] \\ p^{(\frac{1}{2})} &= -\frac{\gamma M^2 v_{\frac{1}{2}}}{\beta}\left[\frac{1 - A_{\frac{1}{2}}}{(x - \beta y)^{\frac{1}{2}}} - \frac{A_{\frac{1}{2}}}{(x + \beta y)^{\frac{1}{2}}}\right] \end{aligned}\right\} \tag{7.4}$$

where the constant $A_{\frac{1}{2}}$ is defined by

$$A_{\frac{1}{2}} = \frac{1 - \beta \tan \lambda}{(1 - \beta \tan \lambda) + \left[\dfrac{1 - \beta \tan \tau_0}{1 + \beta \tan \tau_0}\right]^{\frac{1}{2}}(1 + \beta \tan \lambda)} \tag{7.5}$$

Similarly, the quantities $\tau_{\frac{1}{2}}$, $S_{\frac{1}{2}}$, etc. can be deduced.

8. Discussion of the results. The solutions given in (3.3) and (3.10) show that the first-order solutions such as pressure, velocity, etc. due to reflected and incident waves bear a common ratio $A_1\sqrt{x - \beta y}/(1 - A_1)\sqrt{x + \beta y}$. As $x - \beta y$ is less than $x + \beta y$ in the inviscid field, this ratio will be everywhere less than the value $A_1/(1 - A_1)$ at the surface $y = 0$. According to Fig. 2, this value can be at most 0.14 for $\theta = 15°$ at a flight Mach number of 12. Therefore, for θ less than 10°, the effects due to reflected waves are generally of no consequence and may be neglected.

It is interesting to point out that although the deflection associated with the reflected waves is positive, the pressure and temperature they induce are both negative. In this respect, they act in the opposite sense to that of the entropy wake, as shown below. Hence, in speaking of the effects of vorticity, these distinctions have to be made.

As to the entropy effects, it is noted, first of all, that, following the solutions

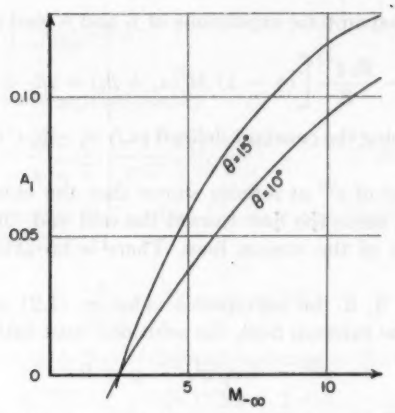


FIG. 2. Variation of A_1 with $M_{-\infty}$ for $\theta = 10°$, $15°$, and $\gamma = 1.40$

of the preceding section, the pressure due to the entropy is of the order $\epsilon^{\frac{1}{2}}$. With this correction, the total pressure, for $M \gg 1$, is

$$p - 1 = 0.596\gamma(\gamma - 1)(1 - 2A_1)\chi\left[1 - \frac{\alpha_{\frac{1}{2}} B_1(1 - 2A_{\frac{1}{2}})}{4.77(1 - 2A_1)} \sqrt[4]{\frac{C}{Re_x}} + \cdots\right]$$

with

$$\chi = M^3\sqrt{C/Re_x} \tag{8.1}$$

This shows in the first place that the pressure correction due to the first-order entropy generated by the perturbed shock is negative, and, moreover, by numerical calculation it is found to be small. From Figs. 3 and 4, at flight Mach 10 and $\theta = 15°$, the correction amounts to only 6 per cent. This is in direct contradiction to the prediction of Professor Lees (Ref. 9) that shock heating due to the generation of entropy induces a non-trivial positive pressure.

Furthermore, the velocity reduction by entropy is also small, from (6.1) and (6.12) it can be verified that

$$u = \tfrac{1}{2}f_0' - \frac{B_1}{2} f_{\frac{1}{2}}' \sqrt[4]{\frac{C}{Re_x}} + \cdots \tag{8.2}$$

From Fig. 3, B_1 is less than unity and, for $C/Re_x < 1$, the correction should be everywhere small and in fact, by Fig. 5, can be at most 10 per cent. The local skin friction is shown to be

$$C_f = \frac{p^{\frac{1}{2}}}{2}\sqrt{\frac{C}{Re_x}}\left[\alpha_0 - B_1\alpha_{\frac{1}{2}}\sqrt[4]{\frac{C}{Re_x}} + \cdots\right] \tag{8.3}$$

and the correction in this case is also small as long as $C/Re_x < 1$.

While it is true that the effect of entropy on pressure and velocity (u) is negligible, it contributes a considerable portion to the temperature near the

edge of the boundary layer. Under the conditions specified in Fig. 6, at places the correction could be 100 per cent at the edge of the boundary layer. It can therefore be concluded that the shock heating of the gas in the cool layer is absorbed largely by expansion of the flow under practically the pressure $\epsilon p^{(1)}$ due to viscous deflection alone.

From the velocity and temperature distributions shown in Figs. 5 and 6, it is observed that as the leading edge is approached, the boundary layer deviates gradually from the unperturbed one and becomes less and less distinct from the inviscid field. This probably is generally the case whenever the external flow is rotational and has a large velocity gradient. Under this circumstance, as already

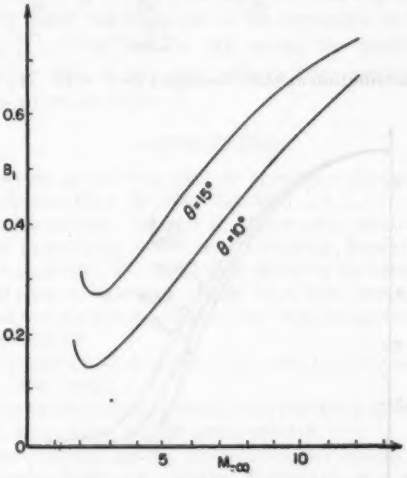


FIG. 3. Variation of B_1 with $M_{-\infty}$ for $\theta = 10°$, 15° and $\gamma = 1.4$

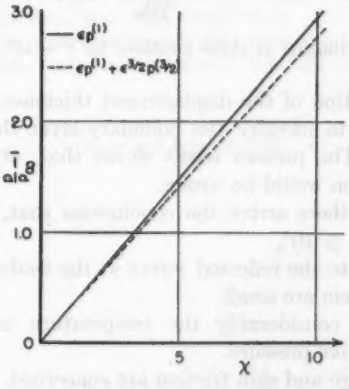


FIG. 4. Surface pressure for $\theta = 15°$, $M_\infty = 5.3$ and $\gamma = 1.40$

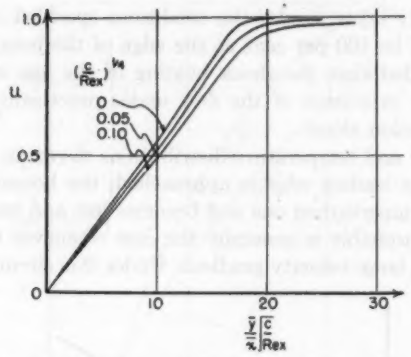


FIG. 5. Velocity distribution at three locations for $\theta = 15°$, $M_\infty = 5.3$ and $\gamma = 1.4$

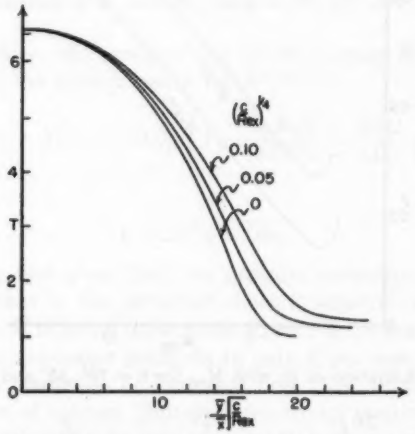


FIG. 6. Temperature distribution at three locations for $\theta = 15°$, $M_\infty = 5.3$ and $\gamma = 1.40$

pointed out, the definition of the displacement thickness is ambiguous, and it would be meaningless to identify the boundary-layer thickness with the displacement thickness. The present result shows that, at least in the case of wedge, this identification would be wrong.

From these results, there arrive the conclusions that, at hypersonic speeds and half wedge angle $\theta \leqq 10°$,

(i) The effects due to the reflected waves at the leading-edge shock and the entropy generated therein are small.

(ii) Entropy raises considerably the temperature in the cool layer but generates a small negative pressure.

(iii) As far as pressure and skin friction are concerned, the application of the simple-wave theory to such problems can be justified.

It is remarked, finally, that, owing to the presence of the singular term S_1, there appear in high-order approximations sequences of singular solutions of which the leading one is given in Section 4. Similarly, the corresponding term $S_{\frac{1}{2}}$ in the $(\frac{1}{2})^{th}$-order solution leads to other sequences of singular solutions of which the leading one is of the order $\epsilon^{\frac{1}{2}}$, and so on. Generally, it is clear that between two consecutive integral powers of ϵ, such as (0, 1), (2, 3) etc., in the expansion of φ there will be a series $\sum_1^\infty \epsilon^{\lambda_n}\varphi^{(\lambda_n)}$, λ_n being a function of the integer n. For example, between $\varphi^{(0)}$ and $\varphi^{(1)}$, the function λ_n is $(2^n - 1)/2^n$. Therefore, if a consistent first-order solution is required, this series has to be calculated. This shows that the method of perturbation as expressed in (6.1) is actually impossible. As long as the vorticity in the external field is singular at $y = 0$, the boundary-layer solution has to be expanded in a sum of two series $\sum_0^\infty \epsilon^n\varphi^{(n)}$ and $\sum_m \sum_n \epsilon^{\lambda_{mn}}\varphi^{(\lambda_{mn})}$. In this sense, the problem cannot be considered as solved—but this in no way affects the conclusions drawn above, which are consistent up to the order $\epsilon^{\frac{1}{2}}$.

REFERENCES

1. Shen, S. F. *An estimate of viscosity effect on hypersonic flow over an insulated wedge*, Jour. of Math. and Phys. **31** (1952) 192–205.
2. Lees, Lester and Probstein, Ronald F., *Hypersonic viscious flow over flat plate*, Aeronautical Engineering Laboratory, Princeton, Report no. 195 (1952).
3. Li, Ting-Yi, and Nagamatsu, H., *Shock wave effects on the laminar skin friction of an insulated flat plate at hypersonic speeds*, Jour. Aero. Sci. **20** (1953) 345–355.
4. Kuo, Y. H., *Viscous flow along a flat plate moving at high supersonic speeds*, Jour. Aero. Sci. **23** (1956) 125–136.
5. Eggers, A. J. and Syvertson, C. A., *Inviscid flow about airfoils at high supersonic speeds*, NACA T.N. 2646 (1952).
6. Chu, Boa-Teh, *On weak interaction of strong shock and Mach waves generated downstream of the shock*, Jour. Aero. Sci. **19** (1952) 433–445.
7. Chapman, D. R. and Rubesin, M. W., *Temperature and velocity profiles in compressible laminar boundary layer with arbitrary distribution of surface temperature*. Jour. Aero. Sci. **16** (1949) 547–565.
8. Howarth, L., *Concerning the effect of compressibility on laminar boundary layers and their separation*, Proc. Roy. Soc., A194 (1948) 16–42.
9. Lees, Lester, *Influence of the leading-edge shock wave on the laminar boundary layer at hypersonic speeds*, Technical Report no. 1, GALCIT (1954).

Cornell University

(Received July 26, 1955)

RAYLEIGH'S PROBLEM AT LOW MACH NUMBER ACCORDING TO THE KINETIC THEORY OF GASES*

By Hsun-Tiao Yang and Lester Lees

I. Introduction. The problem of an infinite flat plate set impulsively into uniform motion in its own plane in an "incompressible" viscous fluid was first treated by Rayleigh.[1] For a compressible fluid the problem has been examined for "small" times by Howarth[2], for "large" times by Van Dyke[3], and, recently, for "intermediate" times by Stewartson[4]. (Here the time is measured in terms of the parameter a^2t/ν.)* Howarth[2] questioned the significance of his solution because of the extremely short time interval during which the pressure on the plate surface decays to a fraction of its value immediately after the start of the motion. For standard conditions of initial temperature and pressure, the plate pressure falls to one-half of its maximum value in 5×10^{-10} seconds, and to one-tenth of this value in 10^{-8} seconds, as compared with an average time τ between successive collisions of a molecule of about 10^{-10} seconds. Over such short time intervals there are only a few collisions per molecule, and the gas can no longer be treated on the basis of the Navier-Stokes equations. In fact, for $t/\tau = 0(1)$ one would expect that the situation corresponds more closely to "free-molecule" flow, where the collisions between the gas molecules and the plate surface are much more important than the collisions between molecules in the gas.

These considerations suggest that Rayleigh's problem should be attacked by means of the kinetic theory of gases in order to determine the initial history of the important physical quantities and the manner in which these quantities approach the behavior predicted by the Navier-Stokes equations for $t/\tau \gg 1$. In this way the whole time-history is treated within a single theoretical framework. Of course, the present problem is highly idealized because of the absence of a leading edge, but the results obtained for the well-defined Rayleigh problem should provide a qualitative understanding of molecular effects in more complex cases, just as Rayleigh's original solution contributed to the understanding of boundary layer flows and the concept of similarity at high Reynolds numbers. For example, a certain analogy exists between Rayleigh's problem and the development of the steady flow downstream of the leading edge of a thin, flat plate. One of the eventual tasks of the present study is to clarify this analogy and to distinguish carefully between molecular effects and viscous effects that would ordinarily be expected at low local Reynolds numbers within the Navier-Stokes framework.

Previous studies of molecular effects in Rayleigh's problem have dealt mainly with the "first-order" influence of velocity-slip and temperature-jump at the

* The quantity $(a^2t/\nu)^{\frac{1}{2}}$ is proportional to the ratio of the speed of sound to the rate of diffusion of vorticity. Alternatively, the parameter $a^2t/\nu = Re/M^2$ is proportional to the ratio of the elapsed time from the start of the motion to the average time between successive collisions of a molecule, $\tau(Re/M_0^2 = \frac{4}{5}t/\tau)$.

plate surface. The Rayleigh problem in slip-flow was first investigated by Schaaf[6], who utilized the classical heat conduction equation governing the velocity component parallel to the plate for a fluid of constant properties. Instead of the usual boundary condition that the gas moves with the plate, Schaaf imposed the "slip" boundary condition. He solved this system for the tangential velocity component and obtained the skin friction by differentiation. The Rayleigh problem was then related to the steady boundary layer flow over a semi-infinite plate by the transformation $x = KUt$, where x is the distance from the leading edge along the plate in the boundary layer flow, U and t are, respectively, the impulsive plate velocity and time in the unsteady Rayleigh problem, and K is a numerical constant. Although Schaaf's analysis is valid only in the slip-flow regime, he nevertheless extrapolated his results over the whole range. For free molecule flow he obtains a finite skin friction twice the correct value. The same remark applies also to the slip velocity, which would be equal to the plate velocity when extrapolated from Schaaf's result to free molecule flow, while the correct value is half of the plate velocity for a diffusively-reflecting surface. Since the momentum transferred from the plate to the gas differs by a factor of two in the two cases, so does the skin friction coefficient.

The trend of the variation of skin friction coefficient with the parameter $\sqrt{Re}/M$, predicted by Schaaf's theory, is followed surprisingly well by the experiments of Sherman[6], who measured the drag force on a number of plates in steady flow under the following conditions: (1) Mach number between 2.3 and 2.9, and Reynolds number between 30 and 800 per inch; (2) Mach number between 2.7 and 3.6, and Reynolds number between 55 and 1850 per inch; and (3) Mach numbers of 0.2 and 0.6, and Reynolds number between 15 and 480 per inch. These cover a range of the parameter $0.008 < M/\sqrt{Re} < 0.38$. However, no direct comparison between theory and experiment can be made because Schaaf's analysis does not take into account the effects of the interaction between the shock wave and the viscous layer in supersonic flow, and the effect of the finite length of the plate.

Mirels[7] extended Schaaf's work to the case of variable fluid properties, but his work seems less satisfactory than that of Schaaf. If the impulsive velocity of the plate in Rayleigh's problem is large enough to warrant consideration of compressibility, then the conventional Navier-Stokes equations would be inadequate in describing the problem in slip-flow.

An analysis similar to Schaaf's was made by Kane[8] for the heat transfer from a flat plate with a temperature-jump boundary condition. Since the energy equation is linearized, Kane's result would also be valid only for low Mach number flow. This heat transfer analysis was extended later to cylinders, spheres[9], and cones[10].

By dealing directly with the steady, uniform flow over a semi-infinite flat plate, Maslen[11] obtained the second approximation to the laminar boundary layer in slip flow. He assumed that the velocities and the thermodynamic properties of the fluid may be expanded in powers of the small quantity $\epsilon = (M\sqrt{\gamma})/(\sqrt{Re})$, which is essentially the ratio of the molecular mean free

path to the boundary layer thickness. By introducing these expansions into the Burnett equations and also into the velocity-slip and the temperature-jump boundary conditions, one finds that the conventional boundary layer equations with velocity slip and temperature jump boundary conditions are correct to the order of ϵ. The Burnett terms give contributions of the order of ϵ^2 and higher. The second approximation, correct to the order of ϵ, indicates a decrease in heat transfer and for supersonic flow an increase in skin friction. For subsonic flow there is no first order shear effect. The change in heat transfer is caused by the slip, but the effect on the skin friction is connected with the interaction between the boundary layer and the external flow. Maslen also obtained the incompressible Rayleigh solution (Appendix D, Reference 11) and found the same subsonic features: namely, zero first-order shear and a decrease in heat transfer.

Independently of Maslen, Nonweiler[12] also investigated the laminar boundary layer in slip flow at low Mach numbers. He found a skin friction increase due to slip as a product of dp/dx and $M/\sqrt{Re}$. If the pressure gradient dp/dx is zero, then slip has no first order shear effect. The effect of slip alone is solely responsible for a heat transfer decrease. Nonweiler also found no effect of temperature jump on skin friction in the cases he studied.

All the results discussed above, as restricted by the framework used, are applicable only to the border regime between continuum and slip flow and could not be expected to hold over the whole domain. Two approaches have been employed to investigate molecular effects in the flow over a flat plate: namely, the study of the Rayleigh problem and the study of the steady boundary layer flow. In the latter case, the problem is not well-defined because of the presence of the leading edge, near which the molecular effects are most important. At any rate, one thing is clear: the ratio $M/\sqrt{Re}$ is the fundamentally-important flow parameter. However, in the boundary layer approach the assumption that the various quantities could be expanded in powers of this parameter is questionable, and the conclusion that the velocity slip has no immediate effect on the skin friction is rather unsatisfactory.

Near the start of the Rayleigh motion, as mentioned before, one expects the flow to resemble a "free-molecule" flow. In the usual free-molecule flow analysis the incident molecules are supposed to be characterized by the Maxwell distribution function based on conditions "far" from the surface. In Rayleigh's problem this distribution function will be modified as soon as a sufficient number of the first reflected molecules have had time to collide with the new incident molecules. Therefore, the behavior of the flow quantities in the initial stages should be described by means of a suitable expansion procedure in terms of the parameter $t/\tau \sim at/L \sim Re/M^2$. Apparently the idea of improving upon the free molecule flow values by seeking solutions of the Maxwell-Boltzmann equation in terms of power series in $1/L$ originated with Jaffé[13]. Keller[14] applied Jaffé's method to the problem of determining the drag on a moving body as well as to the problem of free expansion, but only the first approximation is given explicitly. The general expressions thus obtained for the drag on a convex body in the free molecule flow are the same as those obtained by Heinemann[15] and others. Heine-

mann followed the physical approach of obtaining the drag on arbitrary convex bodies in free-molecule flow by considering the momentum imparted by the gas molecules to the body. The usual assumption of monatomic gas and Maxwellian molecular velocity distribution far from the body was made, and formulas for the drag of a flat plate, a sphere, a cylinder, and an ellipsoid were given. For the particular case of a flat plate moving perpendicular to its plane, Heinemann also obtained a second approximation for the surface pressure to take account of those molecules that collide once in the gas.

It will be interesting to see if solutions of Rayleigh's problem obtained in the present study take the form of series expansions in the parameter t/τ for small times. Conceivably such solutions developed from the free molecule flow regime, combined with the *proper* solutions for large time, may cover most of the range of interest.

Perhaps the most comprehensive attempt to obtain a general approximate solution of the Maxwell-Boltzmann equation is that of H. Grad[16]. In Grad's method the molecular distribution function $f(\vec{\xi}, \vec{x}, t)$ is expanded in terms of Hermite polynomials depending only on $\vec{\xi}$, where $\vec{\xi}$ is the molecular velocity vector, $\vec{x}$ is the space vector, and t is the time. The first term of this series expansion is the Maxwell distribution function based on *local* quantities. The coefficients of the higher terms are taken as the new state variables, which are functions of $\vec{x}$, t, and these coefficients are identified with the "moments" of the distribution function with respect to $\vec{\xi}$. The first moment gives the flux density of molecules, while the higher moments are essentially the normal and shear stresses and the heat flux quantities, which are now regarded as separate dependent variables, not explicitly related to the derivatives of flow velocity $\vec{u}$ and temperature T with respect to $\vec{x}$. By stopping the expansion after a finite number of terms and substituting the Hermite polynomial series for $f(\vec{\xi}, \vec{x}, t)$ into the Maxwell-Boltzmann equation, a system of partial differential equations for pressure, density, velocity, stresses, and heat flux quantities of the gas is obtained. The usual equations expressing conservation of mass, momentum, and energy are always included in the general set. This system of equations should govern all possible flow regimes from free molecule flow to the inviscid or Euler flow.

Grad's boundary conditions at a solid surface are derived by introducing a single "reflection parameter", α, which is similar to the Maxwell-Smoluchowski concepts of reflection coefficient and energy accommodation coefficient. This parameter denotes the fraction of molecules that is reflected specularly. Considering the special case of a perfectly reflecting wall for which $\alpha = 1$, Grad found three of the moments are required to vanish. By working on these three moments for the general case of $0 \leqq \alpha \leqq 1$, he obtained three boundary conditions. However, Grad's mathematical argument of deriving these boundary conditions may be interpreted physically in terms of the conservation laws of mass, momentum, and energy. These boundary conditions are therefore re-derived in Section II.B from a physical point of view.

Because of the uncertainty over the actual physical reflection process[17], it is thought sufficient at the present time to take the tangential momentum reflec-

tion coefficient σ and the energy accommodation coefficient α' as identical and the normal momentum reflection coefficient σ' introduced by Schaaf and Bell[18] as unity. These assumptions are equivalent to Grad's introduction of a single reflection parameter α. Once a sound theoretical foundation is established, it might be possible to obtain α' and σ indirectly by independent measurements of surface pressure and heat transfer rate (or surface temperature), just as the ordinary coefficient of viscosity is obtained from the flow rate in Poiseuille motion.

Grad's method has been applied to the plane shock wave[19] and to the classical problem of the steady flow between two concentric rotating cylinders[20]. In Grad's work[19] on the steady plane shock wave, it was found that the solution breaks down for shocks stronger than $M = 1.65$, which could be taken as the limit of validity of Grad's thirteen-moment equations. The Navier-Stokes solution for the shock profile begins to deviate from the thirteen-moment result at about $M = 1.2$. According to Grad, then, this point of $M = 1.2$ shock strength marks the limit of applicability of the Navier-Stokes equations, although formally the shock solution exists for infinite strength shock in the Navier-Stokes framework.* On the other hand, Gilbarg and Paolucci[21] concluded there is nothing in the shock wave study so far to suggest that the classical continuum theory is not fully adequate to describe the flow in this case, and the recent experiments of Sherman[22] bear them out. However, the question of which system best describes the structure of a shock wave is still in dispute and has to be settled by further research, both theoretical and experimental.

From the above considerations, it is safe to say: (1) Grad's thirteen-moment equations are sufficiently flexible to describe most flows in which there is no shock wave stronger than that of normal oncoming Mach number $M = 1.65$. This last difficulty is certainly avoided in the present study; (2) Although the Navier-Stokes framework yields a formal solution of the shock wave problem even for infinitely strong shocks, the Navier-Stokes' solution for Rayleigh's problem breaks down completely in the initial stage of the flow, giving infinite skin friction. Therefore, the classical equations of Navier-Stokes are definitely not applicable to the study of the initial phases of Rayleigh's problem.

In the present investigation, Grad's thirteen-moment method is applied to Rayleigh's problem. The boundary conditions are restated for the case of a moving surface, and the equivalent assumptions $\sigma = \alpha'$ and $\sigma' = 1$ are retained. At this stage the original restriction to a monatomic gas is thought not to be too severe. In general, the problem is highly non-linear; therefore, this phase of the study deals with the case in which the plate velocity is small compared with the speed of sound in the gas before the start of the motion. The differential equations and boundary conditions are linearized, and momentum and energy are uncoupled. In other words, the problem is purposely simplified to the point where the molecular effects associated with a purely-tangential motion should be clearly discernible. Solutions obtained in this case should provide a basis for attacking Rayleigh's problem with arbitrary plate velocity, where the pressure,

* Of course, relaxation effects would have to be taken into account.

temperature, and velocity fields are interrelated and the problem is more complicated.

II. Equations of Motion and Boundary Conditions. *A. Maxwell-Boltzmann Equation and Grad's Method of Solution.** The basic equation of the kinetic theory of gases is the well-known Maxwell-Boltzmann integro-differential equation for the molecular distribution function $f(\vec{\xi}, \vec{x}, t)$, where $f(\vec{\xi}, \vec{x}, t)\, d\vec{x}\, d\vec{\xi}$ is the mass density of molecules in the physical volume $d\vec{x}$ around $\vec{x}$ having velocities within $d\vec{\xi}$ of $\vec{\xi}$. A new method of solving the Maxwell-Boltzmann equation has been developed by Grad[16]. The method is supposed to give more general solutions than that of Hilbert, Enskog, and Chapman. The essential feature of Grad's method is that the departure of the distribution function from Maxwellian is taken to be linearly dependent on stresses and heat fluxes, which are not assumed to depend explicitly on temperature, pressure, mean velocity, and their gradients. For a monatomic gas sufficiently close to equilibrium, a simple expansion of the distribution function is assumed, of the form

$$f = f^{(0)} \left\{ 1 + \frac{p_{ij}}{2pRT} C_i C_j - \frac{q_i C_i}{pRT} \left(1 - \frac{C^2}{5RT} \right) \right\} \tag{2.1}$$

where

$$f^{(0)} = \frac{\rho}{(2\pi RT)^{\frac{3}{2}}} e^{-C^2/2RT} \tag{2.2}$$

$= $ local Maxwellian distribution, not exactly the equilibrium distribution, since ρ, T, and $\vec{c}$ can be functions of space $\vec{x}$, and time t.

$\rho = $ density of gas

$T = $ temperature of gas

$p = $ pressure of gas

$R = $ gas constant

$C_i = \xi_i - u_i$

$= $ intrinsic molecular velocity

$\xi_i = $ component of molecular velocity in i direction

$u_i = $ component of flow velocity in i direction

$p_{ij} = $ stress tensor (increment over hydrostatic pressure)

$q_i = $ heat flux

It is interesting to compare expansion (2.1) with Enskog's solution to the Maxwell-Boltzmann equation for slowly-varying flow[23]

$$f = f^{(0)} \{ 1 + B_0 + mB_i \xi_i + mB_4 C^2 + \psi_1 + \psi_2 + \psi_3 + \psi_4 \} \tag{2.3}$$

where $B_0, B_1, \cdots, B_4$ are constants, m is mass of a gas molecule and $\psi_1, \cdots, \psi_4$ are functions of ξ_i, C_i, and also of T, ρ (or p), u_i and their gradients. The time dependence of Enskog's solution (2.3) is through that of temperature, density (or pressure), and flow velocity of the gas. However, in Grad's expansion (2.1),

* For details, see Reference 16.

the constituent terms vary with time directly as well as with temperature, pressure, and flow velocity. Arbitrary initial values may be assigned to p_{ij} and q_i and subsequent values determined from the partial differential equations (2.5) below. If one replaces p_{ij} and q_i in expansion (2.1) by the following Navier-Stokes' approximation:

$$p_{ij} = -\mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) + \frac{2}{3}\mu\delta_{ij}\frac{\partial u_k}{\partial x_k}, \qquad q_i = -\lambda\frac{\partial T}{\partial x_i} \tag{2.4}$$

where μ and λ are the coefficients of viscosity and heat conductivity, respectively, then it is seen that Grad's expansion (2.1) will take the form of Enskog's solution.

The expansion (2.1) is valid for all flows in which the macroscopic variables do not vary by a large amount over a mean free path. The variation allowed is sufficient to include the study of shock waves of medium strength, particularly shocks whose normal component of the on-coming Mach number is less than 1.65.* In addition, the basic assumptions for the Boltzmann equation are always imposed.

By substituting expansion (2.1) into the Maxwell-Boltzmann equation, one obtains the Grad equations, including the usual conservation equations and the equation of state

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_r}(\rho u_r) = 0$$

$$\frac{\partial u_i}{\partial t} + u_r\frac{\partial u_i}{\partial x_r} + \frac{1}{\rho}\frac{\partial p}{\partial x_i} + \frac{1}{\rho}\frac{\partial p_{ir}}{\partial x_r} = 0$$

$$\frac{\partial p}{\partial t} + \frac{\partial}{\partial x_r}(u_r p) + \frac{2}{3}p\frac{\partial u_r}{\partial x_r} + \frac{2}{3}p_{ir}\frac{\partial u_i}{\partial x_r} + \frac{2}{3}\frac{\partial q_r}{\partial x_r} = 0$$

$$p = \rho RT$$

$$\frac{\partial p_{ij}}{\partial t} + \frac{\partial}{\partial x_r}(u_r p_{ij}) + \frac{2}{5}\left(\frac{\partial q_i}{\partial x_j} + \frac{\partial q_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\frac{\partial q_r}{\partial x_r}\right) + p_{ir}\frac{\partial u_j}{\partial x_r} + p_{jr}\frac{\partial u_i}{\partial x_r} \tag{2.5}$$

$$- \frac{2}{3}\delta_{ij}p_{rs}\frac{\partial u_r}{\partial x_s} + p\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\frac{\partial u_r}{\partial x_r}\right) + \frac{p}{\mu}p_{ij} = 0$$

$$\frac{\partial q_i}{\partial t} + \frac{\partial}{\partial x_r}(u_r q_i) + \frac{7}{5}q_r\frac{\partial u_i}{\partial x_r} + \frac{2}{5}q_r\frac{\partial u_r}{\partial x_i} + \frac{2}{5}q_i\frac{\partial u_r}{\partial x_r} + RT\frac{\partial p_{ir}}{\partial x_r}$$

$$+ \frac{7}{2}p_{ir}\frac{\partial RT}{\partial x_r} - \frac{p_{ir}}{\rho}\left(\frac{\partial p}{\partial x_r} + \frac{\partial p_{rs}}{\partial x_s}\right) + \frac{5}{2}p\frac{\partial RT}{\partial x_i} + \frac{2}{3}\frac{p}{\mu}q_i = 0$$

where $\delta_{ij} = 0$ when $i \neq j$ and $\delta_{ij} = 1$ when $i = j$, and repeated indices denote summation.

The last two equations above are the same as Eq. (5.18) of Reference 16, except that S_i has been replaced by the more common $2q_i$ and B_1/m by $RT/6\mu$, as in Eq. (5.30) of Reference 16.

* See Introduction, or Reference 19.

In Grad's system, the stresses and the heat fluxes are regarded as new dependent variables governed by additional equations rather than being expressed explicitly in terms of the state variables and their gradients. These additional equations, i.e., the last two of Eq. (2.5), which make the system determinate, may be called, respectively, the stress and the heat equations. It is to be noted that the Euler equations, the Navier-Stokes equations, and the Burnett equations may be deduced from the Grad equations.* Consequently, the Grad equations cover a larger part of the manifold of the solutions to the Boltzmann equation than the others do.

B. Boundary Conditions in Grad's Method. Together with the partial differential equations (2.5), Grad also gives the general boundary conditions. Consider the boundary perpendicular to the x_1-direction. By imposing the condition that a certain fraction α of the incident molecules is specularly reflected and that the remaining molecules are absorbed by the wall and re-emitted with a Maxwellian distribution at the temperature of the wall, T_w, (see Fig. 1) one writes the boundary condition as

$$f^+(\xi_1, \xi_2, \xi_3) = \alpha f^-(-\xi_1, \xi_2, \xi_3) + ke^{-\xi^2/2RT_w} \qquad \xi_1 > 0 \tag{2.6}$$

where

$$f(_w^-) = f^+(\xi) + f^-(\xi)$$

in which

$$f^+(\xi) = 0 \quad \text{for} \quad \xi_1 < 0, \qquad f^-(\xi) = 0 \quad \text{for} \quad \xi_1 > 0$$

and k is determined by the condition that the wall does not collect molecules, i.e.,

$$u_1 = 0 \tag{2.7}$$

The boundary conditions were originally derived by Grad on a mathematical ground of symmetry consideration in the specularly reflected case. However, these boundary conditions may be derived directly from a physical viewpoint based on conservation laws. From expression (2.1) we have the molecular distribution function for the incident molecules,

$$\begin{aligned} f^+(\xi_1, \xi_2, \xi_3) = {} & \frac{\rho}{(2\pi RT)^{\frac{3}{2}}} \exp\left(-\frac{(\xi_1 - u_1)^2 + (\xi_2 - u_2)^2 + \xi_3^2}{2RT}\right) \\ & \times \left[1 + \frac{p_{11}(\xi_1 - u_1)^2}{2pRT} - \frac{p_{11}(\xi_2 - u_2)^2}{2pRT} + \frac{p_{12}(\xi_1 - u_1)(\xi_2 - u_2)}{pRT}\right. \\ & - \frac{q_1(\xi_1 - u_1)}{pRT} - \frac{q_2(\xi_2 - u_2)}{pRT} + \frac{q_1(\xi_1 - u_1)^3}{5p(RT)^2} + \frac{q_1(\xi_1 - u_1)(\xi_2 - u_2)^2}{5p(RT)^2} \\ & \left. + \frac{q_1(\xi_1 - u_1)\xi_3^2}{5p(RT)^2} + \frac{q_2(\xi_1 - u_1)^2(\xi_2 - u_2)}{5p(RT)^2} + \frac{q_2(\xi_2 - u_2)^3}{5p(RT)^2} + \frac{q_2(\xi_2 - u_2)\xi_3^2}{5p(RT)^2}\right] \end{aligned} \tag{2.8}$$

* H. T. Yang: "Reduction of Grad Thirteen-Moment Equations to Burnett Equations for Slip Flow", (unpublished), GALCIT, 1954.

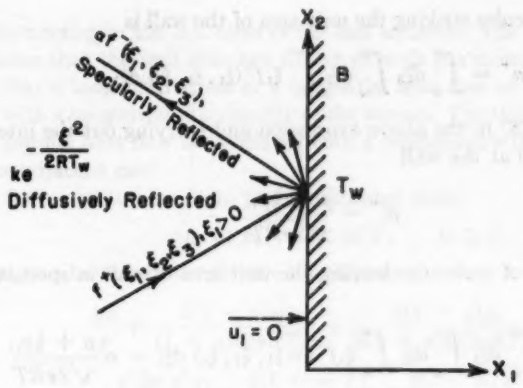


FIGURE 1

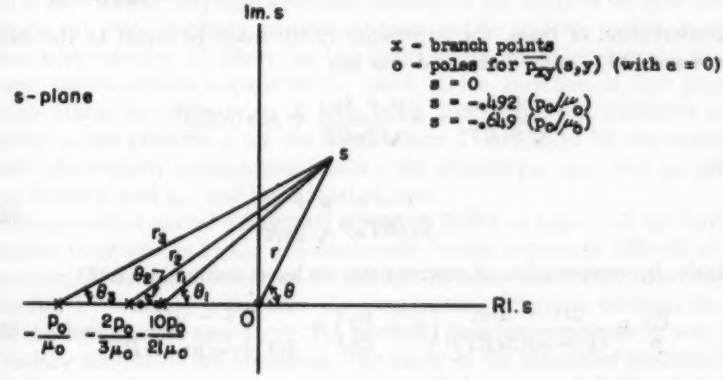


FIGURE 2

FIGS. 1 AND 2

The molecular distribution function for the specularly reflected molecules is obtained by replacing ξ_1 by $-\xi_1$, in expression (2.8). The Maxwellian distribution function at the wall temperature is, by Eq. (2.2),

$$f_w = k \exp\left(-\frac{\xi_1^2 + \xi_2^2 + \xi_3^2}{2RT_w}\right) \tag{2.9}$$

where $k = \rho_w/(2\pi RT_w)^{\frac{1}{2}}$.

The mass of molecules striking the unit area of the wall is

$$m^+ = \int_{-\infty}^{\infty} d\xi_3 \int_{-\infty}^{\infty} d\xi_2 \int_0^{\infty} \xi_1 f^+(\xi_1, \xi_2, \xi_3)\, d\xi_1$$

Substituting Eq. (2.8) in the above expression and carrying out the integration, we get, with $u_1 = 0$ at the wall

$$m^+ = \frac{p + \frac{1}{2}p_{11}}{\sqrt{2\pi RT}} \tag{2.10}$$

Similarly, the mass of molecules leaving the unit area of wall in specular reflection is

$$\alpha m^- = \alpha \int_{-\infty}^{\infty} d\xi_3 \int_{-\infty}^{\infty} d\xi_2 \int_{-\infty}^{\infty} \xi_1 f^-(-\xi_1, \xi_2, \xi_3)\, d\xi_1 = \alpha \frac{p + \frac{1}{2}p_{11}}{\sqrt{2\pi RT}} \tag{2.11}$$

The mass of molecules leaving in Maxwellian reflection is, by Eq. (2.9),

$$\begin{aligned} km_w = k \int_{-\infty}^{\infty} e^{-\xi_3^2/2RT_w}\, d\xi_3 \int_{-\infty}^{\infty} e^{-\xi_2^2/2RT_w}\, d\xi_2 \int_{-\infty}^{0} \xi_1 e^{-\xi_1^2/2RT_w}\, d\xi_1 \\ = k2\pi(RT_w)^2 \end{aligned} \tag{2.12}$$

By conservation of mass, the expression (2.10) must be equal to the sum of expressions (2.11) and (2.12), and one has

$$\frac{p + \frac{1}{2}p_{11}}{\sqrt{2\pi RT}} = \alpha \frac{p + \frac{1}{2}p_{11}}{\sqrt{2\pi RT}} + k2\pi(RT_w)^2$$

Hence,

$$k = \frac{1 - \alpha}{2\pi(RT_w)^2} \frac{p + \frac{1}{2}p_{11}}{\sqrt{2\pi RT}} \tag{2.13}$$

Similarly, by conservation of momentum, we have, using Eq. (2.13)

$$\frac{p_{12}}{p} + \frac{2(1 - \alpha)u_2}{(1 + \alpha)(2\pi RT)^{\frac{1}{2}}} \left(1 + \frac{p_{11}}{2p}\right) + \frac{2(1 - \alpha)q_2}{5(1 + \alpha)p\sqrt{2\pi RT}} = 0 \tag{2.14}$$

By conservation of energy, we have

$$\begin{aligned} \left(\frac{2\pi}{RT}\right)^{\frac{1}{2}} \frac{q_1}{p} + \frac{4(1 - \alpha)}{1 + \alpha} \Bigg[1 - \frac{T_w}{T} + \frac{p_{11}}{2p}\left(\frac{3}{2} - \frac{T_w}{T}\right) \\ - \left(1 + \frac{p_{11}}{2p}\right) \frac{\left(\frac{u_2}{2}\right)^2}{RT} \Bigg] = 0 \end{aligned} \tag{2.15}$$

To sum up, the appropriate boundary conditions for the Grad equations (2.5) are Eqs. (2.7), (2.14), and (2.15). The last two boundary conditions are the same as Eqs. (6.30) and (6.31) in Reference 16, as originally derived by Grad, with S_i replaced by $2q_i$. However, it is to be noted that in Reference 16 the factor

$(\frac{1}{2})^2$ was missing in the last term of the last equation. The first boundary condition states that the wall does not collect or emit gas molecules. The second one states that a tangential stress or a tangential heat flow at the boundary is associated with a tangential slip velocity of the stream. The third one states, roughly, that a normal heat flow is associated with a temperature jump between the wall and the adjacent gas.

For flows not too far from the equilibrium state,

$$p_{11} \ll p, \qquad T_w \cong T, \qquad u \cong 0$$

the last two boundary conditions simplify to

$$\frac{p_{12}}{p} + \frac{2(1-\alpha)u_2}{(1+\alpha)(2\pi RT)^{\frac{1}{2}}} + \frac{2(1-\alpha)q_2}{5(1+\alpha)p(2\pi RT)^{\frac{1}{2}}} = 0 \tag{2.16}$$

$$\left(\frac{2\pi}{RT}\right)^{\frac{1}{2}}\frac{q_1}{p} + \frac{4(1-\alpha)}{1+\alpha}\left(1 - \frac{T_w}{T} + \frac{p_{11}}{4p}\right) = 0 \tag{2.17}$$

which are the boundary conditions (6.30) and (6.31) in Reference 16.

III. Rayleigh's Problem for the Case of Small Impulsive Velocity. *A. Formulation of the Problem.* Rayleigh's problem consists of the study of the flow above an infinite flat plate suddenly accelerated from rest to uniform motion in its own plane with velocity, U. There are two independent variables: namely, the time, t, and the coordinate normal to the plate, y. No variation of flow properties occurs along the plate, viz., $\partial(\ )/\partial x = 0$. The dependent variables are the density ρ, the pressure p (or the temperature T as related by the equation of state), the velocity components, u and v, the stresses p_{xx}, p_{xy}, and p_{yy}, and the heat fluxes q_x and q_y, making a total of nine.

The system of partial differential equations based on Eqs. (2.5) for Rayleigh's problem in general is highly non-linear and, hence, extremely difficult to solve. However, if we limit ourselves to the special case where the plate impulsive velocity, U, is small, and where the temperature difference between the plate and the undisturbed gas, $T_w - T_0$, is small, then the equations as well as the boundary conditions are linearized. The study of the simplified problem should throw some light on the general problem of arbitrary impulsive velocity.

Under the assumptions of small impulsive velocity and small temperature difference, one may write

$$\begin{gathered} u = u', \quad v = v', \quad \rho = \rho_0 + \rho', \quad p = p_0 + p', \quad T = T_0 + T', \\ p_{xx} = p'_{xx}, \quad p_{xy} = p'_{xy}, \quad p_{yy} = p'_{yy}, \quad q_x = q'_x, \quad q_y = q'_y, \end{gathered} \tag{3.1}$$

where the primed quantities denote small perturbations in the undisturbed quantities. The coefficient of viscosity depends on the temperature only:

$$\mu/\mu_0 = (T/T_0)^\omega$$

Hence,

$$\mu = \mu_0\left(1 + \frac{T'}{T_0}\right)^\omega = \mu_0 + u' \tag{3.2}$$

Now substitute expressions (3.1) and (3.2) in Eqs. (2.5). The system of partial differential equations becomes

$$\frac{\partial \rho'}{\partial t} + \rho_0 \frac{\partial v'}{\partial y} = 0 \tag{3.3}$$

$$\frac{\partial u'}{\partial t} + \frac{1}{\rho_0} \frac{\partial p'_{xy}}{\partial y} = 0 \tag{3.4}$$

$$\frac{\partial v'}{\partial t} + \frac{1}{\rho_0} \frac{\partial p'}{\partial y} + \frac{1}{\rho_0} \frac{\partial p'_{yy}}{\partial y} = 0 \tag{3.5}$$

$$\frac{\partial p'}{\partial t} + \frac{5}{3} p_0 \frac{\partial v'}{\partial y} + \frac{2}{3} \frac{\partial q'_y}{\partial y} = 0 \tag{3.6}$$

$$\frac{\partial p'_{xx}}{\partial t} - \frac{2}{3} p_0 \frac{\partial v'}{\partial y} - \frac{4}{15} \frac{\partial q'_y}{\partial y} + \frac{p_0}{\mu_0} p'_{xx} = 0 \tag{3.7}$$

$$\frac{\partial p'_{xy}}{\partial t} + p_0 \frac{\partial u'}{\partial y} + \frac{2}{5} \frac{\partial q'_x}{\partial y} + \frac{p_0}{\mu_0} p'_{xy} = 0 \tag{3.8}$$

$$\frac{\partial p'_{yy}}{\partial t} + \frac{4}{3} p_0 \frac{\partial v'}{\partial y} + \frac{8}{15} \frac{\partial q'_y}{\partial y} + \frac{p_0}{\mu_0} p'_{yy} = 0 \tag{3.9}$$

$$\frac{\partial q'_x}{\partial t} + \frac{p_0}{\rho_0} \frac{\partial p'_{xy}}{\partial y} + \frac{2}{3} \frac{p_0}{\mu_0} q'_x = 0 \tag{3.10}$$

$$\frac{\partial q'_y}{\partial t} + \frac{5}{2} \frac{p_0}{\rho_0} \frac{\partial p'}{\partial y} - \frac{5}{2} \frac{p_0^2}{\rho_0^2} \frac{\partial \rho'}{\partial y} + \frac{p_0}{\rho_0} \frac{\partial p'_{yy}}{\partial y} + \frac{2}{3} \frac{p_0}{\mu_0} q'_y = 0 \tag{3.11}$$

and the equation of state (3.5) becomes

$$p' = R(\rho_0 T' + T_0 \rho') \tag{3.12}$$

or

$$T' = \frac{1}{R\rho_0} p' - \frac{T_0}{\rho_0} \rho' \tag{3.13}$$

The initial conditions are

$$Q'(0, y) = 0 \tag{3.14}$$

where $Q(t, y)$ is any physical quantity. The above conditions state merely that there is no initial disturbance.

The boundary conditions (3.12) to (3.14) are linearized as follows:

$$v'(t, 0) = 0 \tag{3.15}$$

$$\frac{p'_{xy}(t, 0)}{p_0} + \frac{2(1 - \alpha)}{(1 + \alpha)(2\pi R)^{\frac{1}{2}}} \frac{u'(t, 0) - U}{[T_0]^{\frac{1}{2}}} + \frac{2(1 - \alpha)}{5(1 + \alpha)(2\pi R)^{\frac{1}{2}}} \frac{q'_x(t, 0)}{p_0[T_0]^{\frac{1}{2}}} = 0 \tag{3.16}$$

$$\left(\frac{2\pi}{RT_0}\right)^{\frac{1}{2}} \frac{q'_y(t, 0)}{p_0} + \frac{4(1 - \alpha)}{1 + \alpha} \left[1 - \frac{T_w}{T_0} + \frac{p'_{yy}(t, 0)}{4p_0}\right] = 0 \tag{3.17}$$

and these boundary conditions take the same form as (2.16) and (2.17) for flows not too far from equilibrium.

It is also interesting to note that Eqs. (3.4), (3.8), and (3.10) with boundary condition (3.16) completely determine the quantities induced parallel to the plate motion: namely, the tangential flow velocity, u'; the tangential shearing stress p'_{xy}, and the tangential heat flow q'_x. The other quantities like normal flow velocity v', density ρ', pressure p', temperature T', normal stresses p'_{xx}, p'_{yy}, and normal heat flow q'_y are governed by Eqs. (3.3), (3.5), (3.6), (3.7), (3.9), (3.11), and the boundary conditions (3.15) and (3.17). In other words, momentum and energy are not coupled. The dynamic effects of the plate impulsive velocity, U, and the thermal effects of the temperature difference between the plate and the undisturbed gas, $T_w - T_0$, may be treated separately and then superposed.

The characteristics of the linearized system (3.3) to (3.11) are found by the vanishing of the determinant (3.18)

$$
\begin{vmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\rho_0} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{\rho_0} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\rho_0} & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & \frac{5}{3}p_0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{2}{3}p_0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{4}{15} \\
0 & 0 & 0 & 0 & 0 & p_0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \frac{2}{5} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{4}{3}p_0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \frac{8}{15} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{p_0}{\rho_0} & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & -\frac{5}{2}\frac{p_0^2}{\rho_0^2} & 0 & \frac{5}{2}\frac{p_0}{\rho_0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{p_0}{\rho_0} & 0 & 0 & 1 & 0 \\
dt & dy & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & dt & dy & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & dt & dy & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & dt & dy & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & dt & dy & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & dt & dy & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & dt & dy & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & dt & dy & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & dt & dy
\end{vmatrix} = 0
\tag{3.18}
$$

From the above, we obtain

$$\frac{dy}{dt} = 0, \qquad \frac{dy}{dt} = \pm\sqrt{\frac{7}{5}\frac{p_0}{\rho_0}} = \pm\frac{\sqrt{21}}{5}a_0. \tag{3.19}$$

where

$$a_0 = \sqrt{\gamma RT_0} = \sqrt{\tfrac{5}{3}RT_0}$$

is the isentropic speed of sound in the undisturbed monatomic gas, and

$$\begin{aligned} \frac{dy}{dt} &= 0, 0 \\ \frac{dy}{dt} &= \pm\sqrt{\frac{13 \pm \sqrt{94}}{5}\frac{p_0}{\rho_0}} = \pm\frac{\sqrt{3(13 \pm \sqrt{94})}}{5}a_0 \end{aligned} \tag{3.20}$$

The solution $dy/dt = 0$ means that the particle path is one characteristic. The solutions (3.19) and (3.20) represent characteristic directions at "sound-speeds" dy/dt, different from the isentropic sound-speed which is the characteristic slope from the Euler equations. All the characteristics obtained here are identical to those given by Grad in Reference 16, except that due to linearization, the present characteristics are straight lines and are known in advance. It is to be noted that the characteristics (3.19) are associated with the tangential quantities u', p'_{xy}, and q'_x dependent upon the impulsive velocity of the plate U, while the characteristics (3.20) are associated with the rest of quantities dependent upon the temperature difference $T_w - T_0$.

B. Laplace Transform Solutions of the Linearized Rayleigh Problem. To solve the linearized equations (3.3) to (3.11), we apply the Laplace transform defined by[24, 25]

$$\int_0^\infty e^{-st}Q(t, y)\,dt = \bar{Q}(s, y) \tag{3.21}$$

where $Q(t, y)$ is a function defined for all positive values of the variable t, and s is the transform variable

By using Eqs. (3.21) and the initial conditions (3.14), we transform the partial differential equations (3.3) to (3.11) into the following system of ordinary differential equations:

$$\frac{d\bar{v}'}{dy} = -\frac{1}{\rho_0}s\bar{p}' \tag{3.22}$$

$$\frac{d\bar{p}'_{xy}}{dy} = -\rho_0 s\bar{u}' \tag{3.23}$$

$$\frac{d\bar{p}'}{dy} + \frac{d\bar{p}'_{yy}}{dy} = -\rho_0 s\bar{v}' \tag{3.24}$$

$$\frac{5}{3}\frac{d\bar{v}'}{dy} + \frac{2}{3}\frac{1}{p_0}\frac{d\bar{q}'_y}{dy} = -\frac{1}{p_0}s\bar{p}' \tag{3.25}$$

$$\frac{5}{3}\frac{d\bar{v}'}{dy} + \frac{2}{3}\frac{1}{p_0}\frac{d\bar{q}_y'}{dy} = \frac{5}{2}\left(\frac{1}{p_0}s + \frac{1}{\mu_0}\right)\bar{p}_{xx}' \tag{3.26}$$

$$\frac{d\bar{u}'}{dy} + \frac{2}{5}\frac{1}{p_0}\frac{d\bar{q}_x'}{dy} = -\left(\frac{1}{p_0}s + \frac{1}{\mu_0}\right)\bar{p}_{xy}' \tag{3.27}$$

$$\frac{5}{3}\frac{d\bar{v}'}{dy} + \frac{2}{3}\frac{1}{p_0}\frac{d\bar{q}_y'}{dy} = -\frac{5}{4}\left(\frac{1}{p_0}s + \frac{1}{\mu_0}\right)\bar{p}_{yy}' \tag{3.28}$$

$$\frac{d\bar{p}_{xy}'}{dy} = -\left(\frac{\rho_0}{p_0}s + \frac{2}{3}\frac{\rho_0}{\mu_0}\right)\bar{q}_x' \tag{3.29}$$

$$\frac{5}{2}\frac{d\bar{p}'}{dy} - \frac{5}{2}\frac{p_0}{\rho_0}\frac{d\bar{\rho}'}{dy} + \frac{d\bar{p}_{yy}'}{dy} = -\left(\frac{\rho_0}{p_0}s + \frac{2}{3}\frac{\rho_0}{\mu_0}\right)\bar{q}_y' \tag{3.30}$$

The equation of state (3.12) transforms into

$$\bar{T}' = \frac{1}{R\rho_0}\bar{p}' - \frac{T_0}{\rho_0}\bar{\rho}' \tag{3.31}$$

The boundary conditions (3.15) to (3.17) transform to

$$\bar{v}'(s, 0) = 0 \tag{3.32}$$

$$\frac{\bar{p}_{xy}'(s, 0)}{p_0} + \frac{2(1-\alpha)}{(1+\alpha)(2\pi RT_0)^{\frac{1}{2}}}\left[\bar{u}'(s, 0) - U\frac{1}{s} + \frac{1}{5}\frac{\bar{q}_x'(s, 0)}{p_0}\right] = 0 \tag{3.33}$$

$$\left(\frac{2\pi}{RT_0}\right)^{\frac{1}{2}}\frac{\bar{q}_y'(s, 0)}{p_0} + \frac{4(1-\alpha)}{1+\alpha}\left[\left(1 - \frac{T_w}{T_0}\right)\frac{1}{s} + \frac{\bar{p}_{yy}'(s, 0)}{4p_0}\right] = 0 \tag{3.34}$$

From Eqs. (3.23), (3.27), and (3.29) together with boundary condition (3.33), we find the transformed tangential stress, velocity, and heat flow as follows:

$$\bar{p}_{xy}' = p_0\frac{U}{a_0}\frac{1}{\left[\frac{1+\alpha}{1-\alpha}\sqrt{\frac{3\pi}{10}}\,s + \left(\frac{18}{5}s + 2\frac{p_0}{\mu_0}\right)\sqrt{\frac{3s\left(s + \frac{p_0}{\mu_0}\right)}{\left(3s + 2\frac{p_0}{\mu_0}\right)\left(21s + 10\frac{p_0}{\mu_0}\right)}}\right]}e^{-f(s)y} \tag{3.35}$$

$$\bar{u}' = U\frac{1}{\left[\frac{1+\alpha}{1-\alpha}\sqrt{\frac{3\pi}{10}}\,s\sqrt{\frac{s\left(21s + 10\frac{p_0}{\mu_0}\right)}{3\left(s + \frac{p_0}{\mu_0}\right)\left(3s + 2\frac{p_0}{\mu_0}\right)}} + s\left(1 + \frac{3}{5}\frac{s}{3s + 2\frac{p_0}{\mu_0}}\right)\right]}e^{-f(s)y} \tag{3.36}$$

$$\bar{q}_x' = p_0 U \frac{1}{\left[\frac{1+\alpha}{1-\alpha}\sqrt{\frac{\pi}{30}}\, s \sqrt{\frac{\left(21s + 10\frac{p_0}{\mu_0}\right)\left(3s + 2\frac{p_0}{\mu_0}\right)}{3s\left(s + \frac{p_0}{\mu_0}\right)} + \frac{2}{3}\left(\frac{9}{5}s + \frac{p_0}{\mu_0}\right)}\right]} e^{-f(s)y} \tag{3.37}$$

$$f(s) = \frac{5}{a_0}\sqrt{\frac{s\left(s + \frac{p_0}{\mu_0}\right)\left(3s + 2\frac{p_0}{\mu_0}\right)}{3\left(21s + 10\frac{p_0}{\mu_0}\right)}} \tag{3.38}$$

The rest of the transformed quantities, such as normal velocity $\bar{v}'$, density $\bar{\rho}'$, pressure $\bar{p}'$, temperature $\bar{T}'$, normal stresses $\overline{p_{xx}'}$, $\overline{p_{yy}'}$, and normal heat flux $\bar{q}_y'$, are given by Eqs. (A.1) to (A.7) in the Appendix.

It is interesting to note once again from expressions (3.35) to (3.37) and (A.1) to (A.7) that the effect of the impulsive velocity U of the flat plate and the effect of the temperature difference $T_w - T_0$ between the plate and the undisturbed gas are separated by linearization. This separation of effects is made possible mainly because the energy dissipation terms, such as $p_{xy}\,(\partial u/\partial y)$ in the energy equation and $\{u(t) - U\}^2$ in the boundary condition, were neglected in the linearizing procedure. When the impulsive velocity, as well as the temperature difference, is large, the situation would certainly be different.

Our main concern in the present problem is the effect due to impulsive velocity U. Physically, neglecting the temperature difference corresponds to the case of a heat insulated plate, for which the condition $q_{y_{y=0}} = 0$ requires $T_w = T_0$ [Cf. Eq. (A.7) and Eq. (A.11) which gives $B(s) = 0$ for $T_w = T_0$.] In the following sections, Rayleigh's problem of an insulated flat plate at low Mach number will be investigated in detail.

We may foresee the nature of the solutions by examining Eqs. (3.35) to (3.37). For small time, most of the contributions to the quantities p_{xy}, u, and q_x would come from $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x'$, respectively, when s is large. From Eqs. (3.35) to (3.37) we see that when s is large, $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x$ are all of the form

$$\frac{1}{s}\exp\left[-\left(\frac{5}{\sqrt{21}}\right)\left(\frac{y}{a_0}\right)s\right]$$

which gives step-functions in

$$\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right).$$

For large time, the contributions to p_{xy}, u, and q_x would come from $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x$ when s is small, which are of the form

$$\frac{1}{\sqrt{s}}\exp\left[-\left(\frac{y}{\sqrt{\nu_0}}\right)s\right], \quad \frac{1}{s}\exp\left[-\left(\frac{y}{\sqrt{\nu_0}}\right)s\right], \quad \exp\left[-\left(\frac{y}{\sqrt{\nu_0}}\right)s\right],$$

respectively. The corresponding inversion functions are diffusive in character. These distinct features of the solutions for small and large values of the time are just what one would have expected: namely, during the initial stage of the motion, the quantities p_{xy}, u, and q_x are impulsively developed as step functions in y and the region of main influence grows linearly with time; for later times, these quantities are diffused out smoothly in the whole space by viscosity, and the viscous layer grows as $\sqrt{t}$.

IV. The Heat Insulated Plate. For Rayleigh's problem of an insulated infinite flat plate, i.e., $q_y(t, 0) = 0$, we have from expressions (A.1) to (A.11) with $1 - (T_w/T_0) = 0$,

$$p' = 0, \quad p'_{yy} = 0, \quad p'_{zz} = 0, \quad v' = 0, \quad \rho' = 0, \quad T' = 0, \quad q'_y = 0$$

Thus, Eqs. (3.1) become

$$\begin{gathered} u = u', \quad v = 0, \quad \rho = \rho_0, \quad p = p_0, \quad T = T_0 = T_w, \\ p_{xx} = 0, \quad p_{xy} = p'_{xy}, \quad p_{yy} = 0, \quad q_x = q'_x, \quad q_y = 0, \end{gathered} \tag{4.1}$$

i.e., the density, the pressure, and the temperature of the gas remain unchanged; no normal velocity, normal stress, or normal heat flow is developed. The impulsive velocity gives rise to tangential velocity, tangential stress, and tangential energy (heat) flux only.

However, it is important to note that the solutions for the tangential quantities p_{xy}, u, and q_x so obtained apply whether the plate is insulated or not, because of the separability of effects in the linearized problem. For a non-insulated plate, in addition to p_{xy}, u, and q_x, we superpose the other quantities p', p_{yy}, p_{zz}, v, ρ', T', and q_y obtained by inverting Eqs. (A.1) to (A.7).

A. Discussion of the Transformed Quantities in the s-Plane. In the expressions (4.1) the non-vanishing perturbation quantities u, p_{xy}, and q_x are to be obtained by inverting Eqs. (3.35) to (3.37). The shear stress p_{xy}, gas velocity u, tangential energy transfer q_x may be obtained from an inversion integral of the form[24]:

$$Q(t, y) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{st}\bar{Q}(s, y)\, ds \tag{4.2}$$

where c is a constant greater than the real part of all the singularities of $\bar{Q}(s, y)$.

From expressions (3.35) to (3.37), we see that the functions $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x$ all have branch points at

$$s = -\frac{p_0}{\mu_0}, \quad -\frac{2}{3}\frac{p_0}{\mu_0}, \quad -\frac{10}{21}\frac{p_0}{\mu_0}, \quad 0 \tag{4.3}$$

and also poles at roots of each denominator. (For $\alpha = 0$, the roots of the denominator in Eq. (3.35) are 0, $-.492(p_0/\mu_0)$, $-.649(p_0/\mu_0)$.)

All three denominators are essentially positive for real positive s and will not

vanish. This property is demonstrated by representing the complex variable in polar coordinates as

$$s = re^{i\theta} \quad \text{where} \quad |\theta| < \pi/2 \tag{4.4}$$

and the other three factors in the radical accordingly as shown in Fig. 2. The radical in the denominator of the function $\bar{p}_{xy}$ may thus be written as

$$\sqrt{\frac{3s\left(s + \frac{p_0}{\mu_0}\right)}{\left(3s + 2\frac{p_0}{\mu_0}\right)\left(21s + 10\frac{p_0}{\mu_0}\right)}} = \frac{1}{\sqrt{21}}\sqrt{\frac{rr_3}{r_2r_1}}\exp\left[\frac{i}{2}(\theta + \theta_3 - \theta_2 - \theta_1)\right] \tag{4.5}$$

with notations shown in Fig. 2. Now

$$\begin{aligned} |\theta + \theta_3 - \theta_2 - \theta_1| &= |\theta - \theta_1 - (\theta_2 - \theta_3)| \\ &< |\theta - \theta_1| + |\theta_2 - \theta_3| \\ &< |\theta| + |\theta_2| < |\theta| + |\theta| = 2\theta \end{aligned}$$

But $|\theta| < \pi/2$ for s having positive real part

$$\tfrac{1}{2}|\theta + \theta_3 - \theta_2 - \theta_1| < |\theta| < \pi/2$$

Hence the radical (4.5) has positive real part and the denominator of the function $\bar{p}_{xy}$ given by Eq. (3.35) will not vanish for any s with positive real part. In other words, the function $\bar{p}_{xy}$ has no poles in the half-plane $R1\ s > 0$. Likewise, we show the same for $\bar{u}$ and $\bar{q}_x$. To sum up, we have the following statement:

All the three functions $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x$ given by expressions (3.35) to (3.37) have no singularities in the half-plane $R1\ s > 0$. (4.6)

Physically, this must be so; otherwise, the quantities p_{xy}, u, and q_x inverted from Integral (4.2) will diverge with time.

However, we are not going to apply the inversion integral (4.2), since the complete solutions are not only laborious to obtain, but also inconvenient to use. Instead, we shall find solutions suitable for small and large values of the time, which together would cover most of the flow regime, and also approximate solutions which apply to the whole regime.

From expressions (3.35) to (3.37) we see that the function $\bar{p}_{xy}(s, y)$, $\bar{u}(s, y)$, and $\bar{q}_x(s, y)$ have singularities only in the finite s-plane and are regular for s sufficiently large. We may then expand these functions in powers of $1/s$ and invert them term by term to obtain solutions $p_{xy}(t, y)$, $u(t, y)$, and $q_x(t, y)$ for small values of the time as in Heaviside's series expansion. For large values of the time, one expands these functions in powers of s and then inverts.[25]

To obtain approximate solutions for all values of the time, we try to approximate the functions $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x$ of Eqs. (3.35) to (3.37) by simpler functions that can be easily inverted. These simpler functions are chosen in such a way

that the regular behavior at $s = \infty$ and the singular behavior at $s = 0$ are retained, while the singularities with moduli between 0 and ∞ are collapsed into a singularity simpler in nature. By inverting these simpler functions, we are able to obtain approximate solutions which approach the exact solutions for small values and large values of the time and which satisfactorily approximate the exact solutions for intermediate values of the time. The detailed discussions of the different approximations to the transforms will be taken up in the next section.

B. Comparison of Approximations to the Transforms. In this section, the exact expressions (3.35) to (3.37) are compared with their different approximations, which are series expansions in powers of $1/s$, in powers of s, and a simple approximate form for all values of s, corresponding respectively to solutions $p_{xy}(t, y)$, $u(t, y)$, and $q_x(t, y)$ for small values of the time, for large values of the time, and approximate solutions for all values of the time. For simplicity, we consider in detail only the expression (3.35) for $\bar{p}_{xy}$ in the special case of $y = 0$ and $\alpha = 0$, i.e., the shear stress at a diffusively reflecting wall, which interests us most. But the argument applies also to the other two expressions (3.36) and (3.37) as well as to the general case of $y \neq 0$ and $\alpha \neq 0$.

From expression (3.35) with $y = 0$ and $\alpha = 0$, we have

$$\frac{\bar{p}_{xy}(s, 0)}{\mu_0 M_0} = \frac{1}{\sqrt{\frac{3\pi}{10}\frac{\mu_0}{p_0}s + \left(\frac{18}{5}\frac{\mu_0}{p_0}s + 2\right)\sqrt{\frac{3\frac{\mu_0}{p_0}s\left(\frac{\mu_0}{p_0}s + 1\right)}{\left(3\frac{\mu_0}{p_0}s + 2\right)\left(21\frac{\mu_0}{p_0}s + 10\right)}}}} \tag{4.7}$$

The numerical values of $\bar{p}_{xy}(s, 0)/\mu_0 M_0$ vs. $\mu_0/p_0 s$ are computed from Eq. (4.7) and tabulated in Table 1. By expanding the function $\bar{p}_{xy}(s, 0)/\mu_0 M_0$ given by Eq. (4.7) in powers of $1/s$, we have

$$\begin{aligned}\frac{\bar{p}_{xy}(s, 0)}{\mu_0 M_0} &= \frac{1}{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}}\left[\frac{p_0}{\mu_0}\frac{1}{s} - \frac{\frac{61}{21\sqrt{35}}}{\sqrt{\frac{\pi}{2}} + \frac{6}{\sqrt{35}}}\left(\frac{p_0}{\mu_0}\frac{1}{s}\right)^2 \right.\\ &\qquad \left. + \frac{\frac{421}{588}\sqrt{\frac{\pi}{70}} + \frac{11{,}231}{30{,}870}}{\left(\sqrt{\frac{\pi}{2}} + \frac{6}{\sqrt{35}}\right)^2}\left(\frac{p_0}{\mu_0}\frac{1}{s}\right)^3 + \cdots\right] \qquad (4.8)\\ &= 0.569\left[\frac{p_0}{\mu_0}\frac{1}{s} - .247\left(\frac{p_0}{\mu_0}\frac{1}{s}\right)^2 + .0785\left(\frac{p_0}{\mu_0}\frac{1}{s}\right)^3 + \cdots\right]\end{aligned}$$

Eq. (4.8) is an approximation to Eq. (4.7) for large values of s. The numerical calculations are entered in Table 1.

TABLE 1

Comparison of Laplace Transform Solutions for Surface Shear Stress $\frac{\bar{p}_{xy}(s, 0)}{\mu_0 M_0}$

$\frac{\mu_0}{p_0} s$	Original Eq. (4.7)	Large s Eq. (4.8)	Small s Eq. (4.11)	Approximate Eq. (4.12)
0	∞		∞	∞
.0001	127.3		127.5	126.2
.001	39.34		39.25	38.09
.01	11.42		11.45	10.52
.05	4.425		4.37	3.831
.1	2.82	36.30	3.63	2.378
.2	1.742	4.913	4.04	1.433
.4	1.032	1.242	7.53	.839
.5	.863	.933	9.95	.701
.6	.744	.765	13.8	.605
1	.483	.472	26.2	.395
3	.178	.176		.151
5	.1093	.1086		.0951
10	.0558	.0556		.0498
50	.0113	.0113		.0107
100	.0057	.0057		.0054
1,000	.00057	.00057		.00056
10,000	.000057	.000057		.000057
∞	0	0		0

In order to obtain solutions for the shear stress p_{xy} for t/τ or $(a^2t)/\nu \gg 1$, we shall make use of the method of asymptotic expansions given in Reference 25.* According to this approach, an asymptotic expansion of $Q(t)$ for large t can be deduced from the behavior of $\bar{Q}(s)$ near its singularity with the algebraically largest real part. In our case the singularity with largest real part lies at $s = 0$. If $\bar{Q}(s)$ can be expanded near the origin as

$$\bar{Q}(s) = \sum_{n=0}^{\infty} a_n s^{n-1} + \sum_{n=0}^{\infty} b_n s^{n-\frac{1}{2}} \tag{4.9}$$

then $Q(t)$ has the asymptotic expansion for large t

$$Q(t) \sim a_0 + \frac{1}{\pi} \sum_{n=0}^{\infty} (-)^n b_n \Gamma(n + \tfrac{1}{2}) t^{-n-\frac{1}{2}} \tag{4.10}$$

Expanding the function $\bar{p}_{xy}(s, 0)$ given by Eq. (4.7) in the form of Eq. (4.9), we have

$$\frac{\bar{p}_{xy}(s, 0)}{\mu_0 M_0} = \frac{1}{\sqrt{\frac{3}{5}}\sqrt{\frac{\mu_0}{p_0} s}} \left[1 - \sqrt{\frac{\pi}{2}} \sqrt{\frac{\mu_0}{p_0}} s + \left(\frac{\pi}{2} - \frac{1}{2}\right) \frac{\mu_0}{p_0} s \right.$$
$$\left. - \sqrt{\frac{\pi}{2}} \left(\frac{\pi}{2} - 1\right) \left(\frac{\mu_0}{p_0} s\right)^{\frac{3}{2}} + \left(\frac{3571}{200} + \frac{3\pi}{4}\right) \left(\frac{\mu_0}{p_0} s\right)^2 + \cdots \right] \tag{4.11}$$

* For detailed discussion of the conditions under which the expansion is valid and the proof of the general result, see §126, pp. 279–282, Reference 25.

$$= 1.291 \sqrt{\frac{p_0}{\mu_0}} \frac{1}{s} \left[1 - 1.253 \sqrt{\frac{\mu_0}{p_0}} s + 1.071 \frac{\mu_0}{p_0} s \right.$$

$$\left. - .715 \left(\frac{\mu_0}{p_0} s \right)^{\frac{1}{2}} + 20.21 \left(\frac{\mu_0}{p_0} s \right)^2 - \cdots \right]$$

Eq. (4.11) is an approximation to Eq. (4.7) for small values of s. The numerical values are entered in Table 1. From Eq. (4.11) or Table 1, it is clear that the convergence of the series (4.11) becomes poor when $(\mu_0/p_0)\, s \geqq .01$, or $\sqrt{(\mu_0/p_0)s} \geqq .1$, which corresponds to $\sqrt{(p_0/\mu_0)t} = \sqrt{3/5}\ (\sqrt{Re})/M_0 \leqq 0(10)$. In other words, the asymptotic series in the brackets of Eq. (4.11) is useful only in a narrow region around the origin. This result implies that the Burnett-type expansion of the solution for $p_{xy}(t, 0)$ is inappropriate for $\sqrt{Re/M_0} < 10$, or $t/\tau < 100$ (see Section IV.E.1). Therefore, it is necessary to find approximate solutions which will properly bridge the gap between the initial free molecule flow and the later classical Rayleigh motion.

First a simple approximation for $\bar{p}_{xy}\ (s, 0)/\mu_0 M_0$ is required for all values of s. For s large, it is seen from Eq. (4.8) that the dominant term for $\dfrac{\bar{p}_{xy}(s, 0)}{\mu_0 M_0}$ is

$$1 \Big/ \left(\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35} \right) \frac{\mu_0}{p_0} s$$

For s small, Eq. (4.11) shows that the dominant term is $1/\sqrt{3/5}\ \sqrt{(\mu_0/p_0)\, s}$. Therefore, the simplest approximation for all values of s is as follows:

$$\frac{\bar{p}_{xy}(s, 0)}{\mu_0 M_0} = \frac{1}{\left(\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35} \right) \frac{\mu_0}{p_0} s + \sqrt{\frac{3}{5}} \sqrt{\frac{\mu_0}{p_0}} s} \tag{4.12}$$

Of course, the approximation is very good for both small and large values of s. The numerical calculation is carried out and tabulated in Table 1, which is also plotted as Fig. 3.

From Table 1 and Fig. 3, it is clear that the approximation (4.12) to the exact expression (4.7) is satisfactory. Eq. (4.12) gives values lower than those given by Eq. (4.7), but the maximum deviation which occurs in the range $.5 < |\, (u_0/\rho_0)\, s\, | < .6$, is less than 20%. The function $\bar{p}_{xy}(s, 0)/\mu_0 M_0$ given by Eq. (4.7) is regular for s ranging from infinity down to the singularity having largest modulus, which is of the order $|\, (\mu_0/p_0)\, s\, | = 0(1)$, so the expansion (4.8) is useful for $\infty \leqq |\, (\mu_0/p_0)\, s\, | \leqq 1$. For small s, the expansion is valid for a domain between 0 and the singularity having smallest modulus, which is close to the origin, so the range of validity is limited. Based on these observations, we shall obtain solutions for small values of the time by inverting expansion (4.8) and solutions for large values of the time by inverting approximate expression (4.12) to replace the unsatisfactory expansion (4.11).

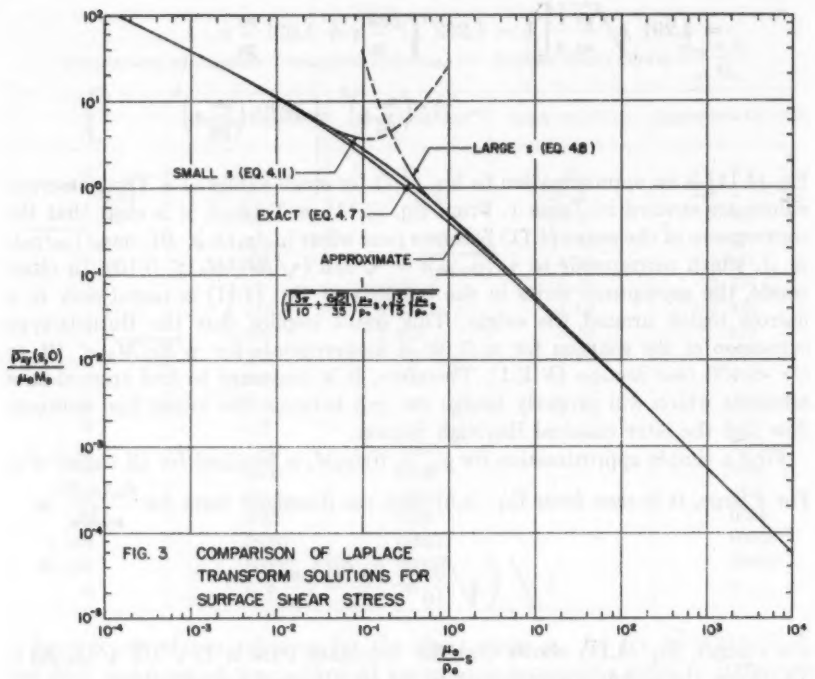

SMALL s (EQ. 4.11)
LARGE s (EQ. 4.8)
EXACT (EQ. 4.7)
APPROXIMATE
FIG. 3 COMPARISON OF LAPLACE TRANSFORM SOLUTIONS FOR SURFACE SHEAR STRESS

FIG. 3

C. Solutions Suitable for Small Values of the Time. By expanding $\bar{p}_{xy}$ given by Eq. (3.35) in powers of $1/s$, we obtain

$$\bar{p}_{xy}(s, y) = p_0 \frac{U}{a_0}\sqrt{\frac{5}{3}} A \exp\left(-\frac{125}{42\sqrt{21}}\frac{p_0}{\mu_0}\frac{y}{a_0}\right)\left[\frac{1}{s} - \left(\frac{61}{21\sqrt{35}} A \frac{p_0}{\mu_0}\right.\right.$$

$$- 0.139\frac{p_0^2}{\mu_0^2}\frac{y}{a_0}\Big)\frac{1}{s^2} + \left\{B\frac{p_0^2}{\mu_0^2} - (0.0681A - 0.0567)\frac{p_0^3}{\mu_0^3}\frac{y}{a}\right. \tag{4.13}$$

$$\left.\left. + 0.00964\frac{p_0}{\mu_0}\frac{y^2}{a_0^2}\right\}\frac{1}{s^3} + 0\left(\frac{1}{s^4}\right)\right]\exp\left(-\frac{5}{\sqrt{21}}\frac{y}{a_0}s\right)$$

where

$$A = \frac{1}{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{\pi}{2}} + \dfrac{6}{\sqrt{35}}}, \qquad B = \frac{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{\pi}{70}}\dfrac{421}{588} + \dfrac{11{,}231}{30{,}870}}{\left(\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{\pi}{2}} + \dfrac{6}{\sqrt{35}}\right)^2}$$

By inverting* and expanding the exponent, we have

* See Table of Transforms, Formulas 61, 62, 63, page 298, Reference 24.

$$p_{xy}(t,y) = p_0 \frac{U}{a_0}\sqrt{\frac{5}{3}}A\left[1 - \frac{p_0}{\mu_0}\left\{\frac{61}{21\sqrt{35}}A\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right) + \frac{125}{42\sqrt{21}}\frac{y}{a_0}\right\}\right.$$
$$+ \frac{p_0^2}{\mu_0^2}\left\{B\frac{1}{2}\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right)^2 + (0.318A + 0.139)\frac{y}{a_0}\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right)\right. \tag{4.14}$$
$$\left.\left. + 0.211\frac{y^2}{a_0^2}\right\}\right] + 0(t^3) \quad t > \frac{5}{\sqrt{21}}\frac{y}{a_0}$$
$$= 0 \qquad 0 < t < \frac{5}{\sqrt{21}}\frac{y}{a_0}$$

where A and B are given in Eq. (4.13).

Similarly, expanding the functions $\bar{u}$, $\bar{q}_x$ given by Eqs. (3.36) and (3.37) in powers of $1/s$ and inverting term by term, we obtain

$$u(t,y) = U\sqrt{\frac{5}{7}}A\left[1 + \frac{p_0}{\mu_0}\left\{C\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right) - \frac{125}{42\sqrt{21}}\frac{y}{a_0}\right\}\right.$$
$$+ \frac{p_0^2}{\mu_0^2}\left\{(C^2 - D)\frac{1}{2}\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right)^2\right.$$
$$\left.\left. + \left(0.139 - \frac{125}{42\sqrt{21}}C\right)\frac{y}{a_0}\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right) + 0.211\frac{y^2}{a_0^2}\right\}\right] \tag{4.15}$$
$$+ 0(t^3) \qquad t > \frac{5}{\sqrt{21}}\frac{y}{a_0}$$
$$= 0 \qquad 0 < t < \frac{5}{\sqrt{21}}\frac{y}{a_0}$$

where A is given in Eq. (4.13) and

$$C = \frac{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{5\pi}{14}}\dfrac{5}{6} + \dfrac{2}{15}}{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{7\pi}{10}} + \dfrac{6}{5}} \qquad D = \frac{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{\pi}{70}}\dfrac{1699}{504} + \dfrac{4}{45}}{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{7\pi}{10}} + \dfrac{6}{5}}$$

$$q_x(t,y) = p_0 U\sqrt{\frac{5}{7}}A\left[1 - \frac{p_0}{\mu_0}\left\{E\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right) + \frac{125}{42\sqrt{21}}\frac{y}{a_0}\right\}\right.$$
$$+ \frac{p_0^2}{\mu_0^2}\left\{(E^2 - F)\tfrac{1}{2}\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right)\right. \tag{4.16}$$
$$\left.\left. + \left(\frac{125}{42\sqrt{21}}E + 0.139\right)\frac{y}{a_0}\left(t - \frac{5}{\sqrt{21}}\frac{y}{a_0}\right) + 0.211\frac{y^2}{a_0^2}\right\}\right]$$
$$+ 0(t^3) \qquad t > \frac{5}{\sqrt{21}}\frac{y}{a_0}$$
$$= 0 \qquad 0 < t < \frac{5}{\sqrt{21}}\frac{y}{a_0}$$

where A is given in Eq. (4.13)

$$E = \frac{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{\pi}{70}}\dfrac{1}{2}+\dfrac{2}{3}}{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{7\pi}{10}}+\dfrac{6}{5}}, \qquad F = \frac{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{\pi}{70}}\dfrac{299}{504}}{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{7\pi}{10}}+\dfrac{6}{5}}$$

For a fixed small value of the time, the distribution of tangential stress, p_{xy}, tangential energy flow, q_x, and tangential gas velocity, u, normal to the plate surface is represented in Figs. 4 and 5, where the numbers 1, 2, and 3 designate the number of terms taken in evaluating the functions.

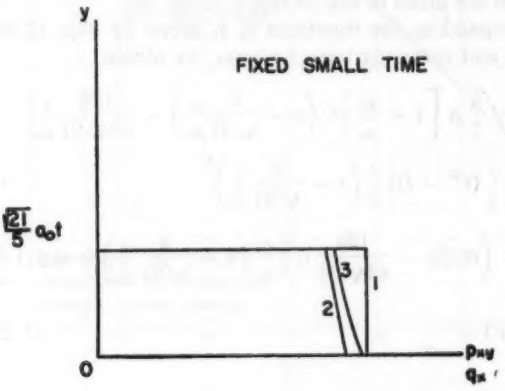


FIGURE 4

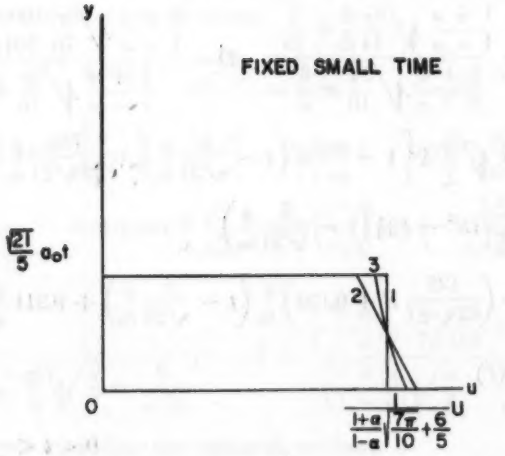


FIGURE 5

FIGS. 4 AND 5

From Eqs. (4.14) to (4.16), the parameter in the small time solution is

$$\frac{p_0 t}{\mu_0} = \frac{3}{5}\frac{\rho_0 U^2 t}{\mu_0}\frac{5}{3}\frac{p_0}{\rho_0}\frac{1}{U^2} = \frac{3}{5}\frac{Re}{M_0^2} \tag{4.17}$$

where

$$Re = \text{instantaneous Reynolds number} = (\rho_0 U^2 t)/\mu_0$$

$$M_0 = \text{Mach number} = U/a_0 = U/\sqrt{(5/3)(p_0/\rho_0)}$$

For small values of the time, or $\rho_0 t/\mu_0 \sim a_0^2 t/\nu_0 \sim t/\tau < 1$, the solution in series expansions of this parameter is indeed satisfactory.

From Figs. 4 and 5, we see that when a heat insulated infinite flat plate is suddenly accelerated into uniform motion with a small impulsive velocity U, the gas far from the plate is not at all disturbed at first, because the gas molecules there are not aware of the plate motion yet. Only the gas close enough to the plate is affected by the motion. Within a thin layer $y < \sqrt{21/5}\, a_0 t$, where $t = 0(\tau)$, i.e., t is of the order of the time between two successive collisions of a gas molecule, the gas as a whole moves with the plate at a fraction of the impulsive velocity U as shown in Fig. 5. Also the tangential stress and the tangential energy flux are developed as shown in Fig. 4. This thin layer grows linearly with time at the rate of $(\sqrt{21}/5)a_0$, which is one of the characteristic slopes dy/dt found in Eq. (3.19).

A discussion of some of the other physical aspects of these solutions is given in Section V.

D. Solutions for Large Values of the Time. The formal solutions for the quantities p_{xy}, u, and q_x at large times are obtained by expanding $\bar{p}_{xy}$, $\bar{u}$, and $\bar{q}_x$ given by Eqs. (3.35) to (3.37) in the form of Eq. (4.9), and writing out the asymptotic solutions in the form of Eq. (4.10). Thus,

$$\begin{aligned}\bar{p}_{xy}(s, y) = p_0 \frac{U}{a_0}\frac{1}{\sqrt{\frac{3}{5}}\sqrt{\frac{p_0}{\mu_0}}}\frac{1}{\sqrt{s}}\Bigg[1 - \left(G\sqrt{\frac{\mu_0}{p_0}} + \frac{y}{\sqrt{\nu_0}}\right)\sqrt{s} + Hs \\ - \left\{(G^3 - G)\left(\frac{\mu_0}{p_0}\right)^{\frac{1}{2}} + \left(G^2 - \frac{3}{10}\right)\frac{\mu_0}{p_0}\frac{y}{\sqrt{\nu_0}}\right. \\ \left. + G\frac{1}{2}\sqrt{\frac{\mu_0}{p_0}}\frac{y^2}{\nu_0} + \frac{1}{6}\frac{y^3}{\nu_0^{\frac{3}{2}}}\right\} s\sqrt{s} + Ks^2 + 0(s^{\frac{5}{2}})\Bigg]\end{aligned} \tag{4.18}$$

where

$$G = \frac{1+\alpha}{1-\alpha}\sqrt{\frac{\pi}{2}} \qquad H = \left(G^2 - \frac{1}{2}\right)\frac{\mu_0}{p_0} + G\sqrt{\frac{\mu_0}{p_0}}\frac{y}{\sqrt{\nu_0}} + \frac{1}{2}\frac{y^2}{\nu_0}$$

$$K = \left(\frac{3}{2}G^2 + \frac{3{,}571}{200}\right)\frac{\mu_0^2}{p_0^2} + \left(G^3 - \frac{4}{5}G\right)\left(\frac{\mu_0}{p_0}\right)^{\frac{1}{2}}\frac{y}{\sqrt{\nu_0}} + \left(\frac{1}{2}G^2 - \frac{1}{20}\right)\frac{\mu_0}{p_0}\frac{y^2}{\nu_0}$$

$$+ G\frac{1}{6}\sqrt{\frac{\mu_0}{p_0}}\frac{y^3}{\nu_0^{\frac{3}{2}}} + \frac{1}{24}\frac{y^4}{\nu_0^2}$$

and

$$p_{sy}(t, y) \sim p_0 \frac{U}{a_0} \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{\frac{3}{5}} \sqrt{\frac{p_0}{\mu_0}} t} \left[1 - \frac{1}{2} H \frac{1}{t} + \frac{3}{4} K \frac{1}{t^2} + 0\left(\frac{1}{t^3}\right)\right] \tag{4.19}$$

where H and K are given in Eq. (4.18). Similarly, we obtain

$$\begin{aligned} u(t, y) \sim U \Bigg\{ 1 - \frac{1}{\sqrt{\pi} t} \Bigg[\left(G \sqrt{\frac{\mu_0}{p_0}} + \frac{y}{\sqrt{\nu_0}}\right) - \Bigg\{ \left(G^2 - \frac{1}{5} G\right) \left(\frac{\mu_0}{p_0}\right)^{\frac{1}{2}} \\ + \frac{1}{2} \left(G^2 - \frac{1}{10}\right) \frac{\mu_0}{p_0} \frac{y}{\sqrt{\nu_0}} + G \sqrt{\frac{\mu_0}{p_0}} \frac{y^2}{\nu_0} + \frac{1}{12} \frac{y^3}{\nu_0^{\frac{3}{2}}} \Bigg\} \frac{1}{t} \\ + \Bigg\{ \left(\frac{153}{200} G - G^3\right) \left(\frac{\mu_0}{p_0}\right)^{\frac{1}{2}} - \left(\frac{33}{40} G^2 + \frac{87}{400}\right) \frac{\mu_0^2}{p_0^2} \frac{y}{\sqrt{\nu_0}} \\ + G^3 \frac{3}{8} \left(\frac{\mu_0}{p_0}\right)^{\frac{1}{2}} \frac{y^2}{\nu_0} + \frac{1}{8} \left(G^2 + \frac{3}{10}\right) \frac{\mu_0}{p_0} \frac{y^3}{\nu_0^{\frac{3}{2}}} + G \frac{1}{32} \sqrt{\frac{\mu_0}{p_0}} \frac{y^4}{\nu_0^2} \\ + \frac{1}{160} \frac{y^5}{\nu_0^{\frac{5}{2}}} \Bigg\} \frac{1}{t^2} + 0\left(\frac{1}{t^3}\right) \Bigg] \Bigg\} \end{aligned} \tag{4.20}$$

where G is given in Eq. (4.18).

$$\begin{aligned} q_s(t, y) \sim p_0 U \frac{3}{4} \frac{1}{\sqrt{2}} \frac{1}{\left(\frac{p_0}{\mu_0} t\right)^{\frac{1}{2}}} \Bigg[\frac{1 + \alpha}{1 - \alpha} + \Bigg\{ \left(\frac{1 + \alpha}{1 - \alpha} \frac{69}{20} - \frac{(1 + \alpha)^3}{(1 - \alpha)^3} \frac{3\pi}{4}\right) \frac{\mu_0}{p_0} \\ + \left(\frac{12}{5} \sqrt{\frac{2}{\pi}} - \frac{1 + \alpha}{1 - \alpha} \sqrt{\frac{\pi}{2}} \frac{3}{2}\right) \sqrt{\frac{\mu_0}{p_0}} \frac{y}{\sqrt{\nu_0}} \\ - \sqrt{\frac{2}{\pi}} \sqrt{\frac{p_0}{\mu_0}} \frac{1}{4} \frac{y^3}{\nu_0^{\frac{3}{2}}} \Bigg\} \frac{1}{t} + 0\left(\frac{1}{t^2}\right) \Bigg] \end{aligned} \tag{4.21}$$

where ν_0 = coefficient of kinematic viscosity = μ_0/ρ_0 .

As we have seen in the previous section the expansions for small s such as Eq. (4.18) are good approximations for $0 < |(\mu_0/p_0)s| < .01$. In other words, solutions (4.19) to (4.21) are convergent only for sufficiently large values of the time, i.e., $t/\tau > 0(100)$. These solutions are included here for comparison with the approximate solutions to be found below.

In the previous section an approximate expression for $\tilde{p}_{sy}(s, 0)$ (Eq. (4.12)) was found by combining the dominant terms for large and small s. Evidently the same procedure can be employed for $\tilde{p}_{sy}(s, y)$, provided that the quantity $f(s)$ in the exponential term $e^{-f(s)y}$ (Eqs. (3.35) and (3.38)) is approximated by the dominant terms

$$-\left(\frac{5}{\sqrt{21}} \frac{s}{a_0} + \frac{\sqrt{s}}{\sqrt{\nu_0}}\right)$$

for large and small s, i.e.,

$$\bar{p}_{xy}(s, y) = p_0 \frac{U}{a_0} \sqrt{\frac{5}{3}} \frac{1}{A^{-1}s + \sqrt{\frac{p_0}{\mu_0}}\, s} \exp\left(- \frac{5}{\sqrt{21}} \frac{y}{a_0} s - \frac{y}{\sqrt{\nu_0}} \sqrt{s} \right) \tag{4.22}$$

where A is given in Eq. (4.13).

By inverting* expression (4.22), we obtain

$$\begin{aligned} p_{xy}(t, y) = {} & p_0 \frac{U}{a_0} \sqrt{\frac{5}{3}} A \exp\left(A \sqrt{\frac{p_0}{\mu_0}} \frac{y}{\sqrt{\nu_0}} \right) \exp\left(A^2 \frac{p_0}{\mu_0} t \right. \\ & \left. \left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t} \right) \right) \times \operatorname{erfc}. \left(A \sqrt{\frac{p_0}{\mu_0} t \left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t} \right)} \right. \\ & \left. + \frac{y}{2\sqrt{\nu_0 t \left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t} \right)}} \right) \qquad t > \frac{5}{\sqrt{21}} \frac{y}{a_0} \\ = {} & 0 \qquad 0 < t < \frac{5}{\sqrt{21}} \frac{y}{a_0} \end{aligned} \tag{4.23}$$

where A is given in Eq. (4.13).

This closed-form solution for the shear stress p_{xy} is not only the proper solution for large values of the time but also the approximate solution for all values of the time. Letting $t \to 0$ and bearing in mind $y = 0(t)$, we have

$$\begin{aligned} \lim_{t \to 0} p_{xy}(t, y) = {} & p_0 \frac{U}{a_0} \sqrt{\frac{5}{3}} A \qquad t > \frac{5}{\sqrt{21}} \frac{y}{a_0} \\ = {} & 0 \qquad 0 < t < \frac{5}{\sqrt{21}} \frac{y}{a_0} \end{aligned} \tag{4.24}$$

This expression for $p_{xy}(t, y)$ is the first term of the small time solution (4.14), which gives the free-molecule shear stress based on Grad's distribution function (2.1). By letting $t \to \infty$ and making use of the asymptotic formula† for the complementary error function

$$\operatorname{erfc}. x \sim \frac{e^{-x^2}}{x\sqrt{\pi}} \left[1 - \frac{2!}{1!(2x)^2} + \frac{4!}{2!(2x)^4} - \frac{6!}{3!(2x)^6} + \cdots \right]$$

we obtain

$$\lim_{t \to \infty} p_{xy}(t, y) = p_0 \frac{U}{a_0} \frac{1}{\sqrt{\frac{3}{5} \pi \frac{p_0}{\mu_0} t}} e^{-y^2/4\nu_0 t}$$

* See Table of Operations, Formula 12, p. 294 and Table of Transforms, Formula 87, p. 300, Reference 24.

† See, for instance, Dwight, H. B.: Table of Integrals, 2nd Ed., p. 129, The Macmillan Company, New York, 1947.

or, by Eq. (4.17)

$$\lim_{t\to\infty} p_{xy}(t, y) = \frac{\rho_0 U^2}{\sqrt{\pi}} \frac{1}{\sqrt{Re}} e^{-y^2/4\nu_0 t} \qquad \text{for all } y \quad (4.25)$$

which is, of course, the classical Rayleigh result.[1]

In other words, the closed-form approximate solution (4.23) gives the exact

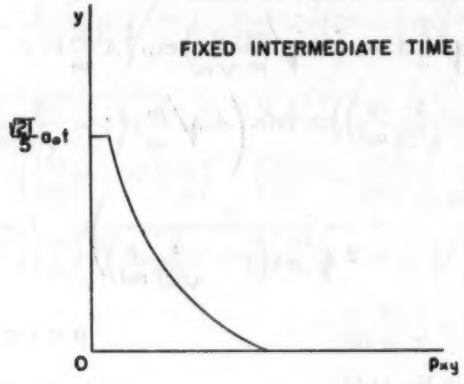


FIGURE 6

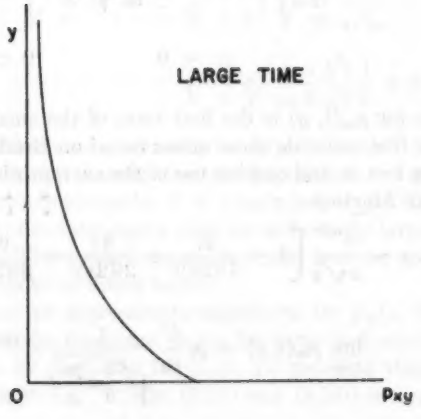


FIGURE 7

FIGS. 6 AND 7

description of the gas behavior for both extremely small and extremely large values of the time. In the initial phase, the flow, impulsively started, is a free-molecule one, and the thin layer grows linearly with time. During the later stages, the flow is diffusive in nature with the viscous layer growing as $\sqrt{t}$ and may be treated as a continuum, as Rayleigh first did. For intermediate values of the time, Eq. (4.23) gives an approximate description of the flow field, which is essentially a diffusive layer bounded by a wave-front travelling at the speed $(\sqrt{21}/5)a_0$. In Fig. 6, the distribution of shear stress p_{xy} with respect to y for a fixed intermediate time is shown. For sufficiently large values of time, for which $y = 0(\sqrt{t})$, we may reduce Eq. (4.23) to

$$p_{xy}(t, y) = p_0 \frac{U}{a_0} \sqrt{\frac{5}{3}} A \exp\left(A \sqrt{\frac{p_0}{\mu_0}} \frac{y}{\sqrt{\nu_0}}\right) \cdot \exp\left(A^2 \frac{p_0}{\mu_0} t\right) \operatorname{erfc}.\left(A \sqrt{\frac{p_0}{\mu_0}} t + \frac{y}{2\sqrt{\nu_0 t}}\right) \tag{4.26}$$

which is again represented in Fig. 7. It is clear that for sufficiently large values of the time, the shear stress distribution is close to Rayleigh's result.

In a similar manner, the proper large time solution for the tangential energy flux q_x is found to be

$$q_x(t, y) = p_0 U \sqrt{\frac{5}{7}} A \frac{1}{a - b} \left[a \exp\left(a \frac{y}{\sqrt{\nu_0}}\right) \exp\left(a^2 t \left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t}\right)\right) \cdot \operatorname{erfc}.\left(a\sqrt{t} + \frac{y}{2\sqrt{\nu_0 t\left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t}\right)}}\right) - b \exp\left(b \frac{y}{\sqrt{\nu_0}}\right) \cdot \exp\left(b^2 t \left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t}\right)\right) \operatorname{erfc}.\left(b\sqrt{t} + \frac{y}{2\sqrt{\nu_0 t\left(1 - \frac{5}{\sqrt{21}} \frac{y}{a_0 t}\right)}}\right) \right] \tag{4.27}$$

$$t > \frac{5}{\sqrt{21}} \frac{y}{a_0}$$

$$= 0 \qquad 0 < t < \frac{5}{\sqrt{21}} \frac{y}{a_0}$$

where

$$\begin{matrix} a \\ b \end{matrix} = \sqrt{\frac{5}{7}} A \sqrt{\frac{p_0}{\mu_0}} \left[\frac{1}{3} G \pm \sqrt{\frac{1}{9} G^2 - \frac{2}{3} \sqrt{\frac{7}{5}} G - \frac{4}{5}} \right]$$

and A and G are given in Eq. (4.13) and Eq. (4.18), respectively. Like Eq. (4.23), Eq. (4.27) is also an approximate solution of the quantity q_x for all values of the time. The discussion of the properties of $p_{xy}(t, y)$ also applies to $q_x(t, y)$.

The situation for the gas velocity $u(t, y)$ is a little different. The dominant terms in the expansions for large and small s are, respectively [Eq. (3.36)]:

$$\bar{u}(s, y) \cong U \frac{1}{\left(\frac{1+\alpha}{1-\alpha}\sqrt{\frac{7\pi}{10}} + \frac{6}{5}\right)s - \left(\frac{1+\alpha}{1-\alpha}\sqrt{\frac{7\pi}{10}}\frac{25}{42} + \frac{2}{15}\right)\frac{p_0}{\mu_0}} e^{-(5/\sqrt{21})(y/a_0)s} \qquad s \text{ large} \tag{4.28}$$

$$\bar{u}(s, y) \cong U \frac{1}{s + \frac{1+\alpha}{1-\alpha}\sqrt{\frac{\pi}{2}}\sqrt{\frac{\mu_0}{p_0}}\, s\sqrt{s}} e^{-(y/\sqrt{\nu_0})\sqrt{s}} \qquad s \text{ small} \tag{4.29}$$

Here the leading terms in both denominators are of the order s, and hence no approximate expression for all s, like Eq. (4.22), can be obtained for the transform $\bar{u}(s, y)$. Therefore, for the velocity of the gas, we shall obtain a closed-form solution valid only for large values of the time rather than for all values of the time. It is obtained by inverting* Eq. (4.29)

$$u(t, y) = U\left[\operatorname{erfc}.\left(\frac{y}{2\sqrt{\nu_0 t}}\right) - \exp\left(G\sqrt{\frac{\rho_0}{p_0}}\frac{p_0}{\mu_0}y\right) \cdot \exp\left(G^2\frac{p_0}{\mu_0}t\right)\operatorname{erfc}.\left(G\sqrt{\frac{p_0}{\mu_0}}t + \frac{y}{2\sqrt{\nu_0 t}}\right)\right] \qquad t \text{ large} \tag{4.30}$$

where G is given in Eq. (4.18).

However, for the special case $y = 0$, it is possible to approximate $U(1/s) - \bar{u}(s, 0)$ for all values of s, since the dominant terms there are of different orders. The velocity slip $U - u(t, 0)$ obtained by inverting $U(1/s) - \bar{u}(s, 0)$ will be given in Section IV.E.2.

E. The Skin Friction, the Slip Velocity, and the Tangential Energy (Heat) Flux of the Gas Adjacent to the Heat Insulated Plate. In this section, we are going to study the quantities which interest us most, viz., the skin friction, the slip velocity, and the tangential energy (heat) flux of the gas adjacent to the plate. All these quantities are obtained by setting $y = 0$ in the expressions for p_{xy}, u, and q_x found in the previous section.

So far the surface condition of the plate is not specified; in other words, the parameter α is rather arbitrary, and it may vary from $\alpha = 1$ for specular reflection to $\alpha = 0$ for diffusive reflection. In the numerical calculation that follows, we consider the case $\alpha = 0$, viz., the gas molecules impinging on the surface of the plate are temporarily absorbed and later re-emitted in a random fashion, such that they give up all their momentum to the plate.

1. The Skin Friction Coefficient. The instantaneous skin friction coefficient is defined as

$$c_f = \frac{p_{xy}(t, 0)}{\frac{1}{2}\rho_0 U^2} \tag{4.31}$$

* See Table of Transforms Formula 86, p. 300, Reference 24.

By putting $y = 0$ and $\alpha = 0$ in Eqs. (4.14), (4.19), and (4.23) and using definitions (4.31) and (4.17), we obtain the following expressions for the skin friction coefficient:

$$M_0 c_f = 0.683\left[1 - 0.0906\frac{Re}{M_0^2} + 0.00878\frac{Re^2}{M_0^4}\right] + 0\left(\frac{Re^3}{M_0^6}\right) \quad t \text{ small} \tag{4.32}$$

$$M_0 c_f = 1.128\frac{M_0}{Re^{\frac{1}{2}}}\left[1 - 0.892\frac{M_0^2}{Re} + 42.1\frac{M_0^4}{Re^2}\right] + 0\left(\frac{M_0^7}{Re^{7/2}}\right) \quad t \text{ large} \tag{4.33}$$

$$M_0 c_f \cong 0.683 \exp\left(0.342\frac{\sqrt{Re}}{M_0}\right) \operatorname{erfc}.\left(0.342\frac{\sqrt{Re}}{M_0}\right) \quad \text{all } t \tag{4.34}$$

The free molecule value of the skin friction coefficient based on Grad's distribution function is obtained by letting $t \to 0$ in either Eq. (4.32) or Eq. (4.33) or by substituting Eq. (4.24) into Eq. (4.31)

$$M_0 c_f = \frac{\frac{6}{5}}{\sqrt{\frac{3\pi}{10} + \frac{6\sqrt{21}}{35}}} = 0.683 \qquad t \to 0 \tag{4.35}$$

which is to be compared with the conventional free molecule skin friction coefficient based on Maxwellian distribution.

$$M_0 c_f = \sqrt{\frac{2}{\pi}\frac{1}{\gamma}} = \sqrt{\frac{2}{\pi}\frac{3}{5}} = 0.618 \tag{4.36}$$

These two values differ because different molecular distribution functions are used (see Section IV.E.2). The Rayleigh result is obtained by letting $t \to \infty$ in either Eq. (4.33) or Eq. (4.34) or by substituting Eq. (4.25) into Eq. (4.31)

$$M_0 c_f = \frac{2}{\sqrt{\pi}}\frac{M_0}{Re^{\frac{1}{2}}} = 1.128\frac{M_0}{Re^{\frac{1}{2}}} \qquad t \to \infty \tag{4.37}$$

The results of the numerical calculation of the skin friction coefficient based on Eqs. (4.32) to (4.34) are tabulated in Table 2 and are also plotted in Fig. 8.

From Fig. 8, we may make the following statements:

(1) The series solution (4.32) in powers of the parameter $(Re/M_0^2) \sim t/\tau$ is indeed quite accurate for small values of the time $(t/\tau \leqq 1)$. As more and more of the reflected molecules collide with the incident molecules the thickness of the layer influenced by the moving plate grows (see also Fig. 4) and the skin-friction decreases with time.

(2) On the other hand, the solution for large time in inverse powers of $(Re/M_0^2) \sim t/\tau$ is of very limited usefulness. When $M_0^2/Re = 0.01$ the second term in the expression for c_f (Eq. (4.33)) is only one per cent of the first term, but the third term is already 40 per cent as large as the second. Clearly, the Burnett-type expansion is inappropriate, at least for Rayleigh's problem.

(3) The approximate solution (4.34) presented here is not only the appropriate solution for large values of the time, but is also a satisfactory solution for

TABLE 2

Time History of Skin Friction Coefficient M_0C_f in Rayleigh's Problem

$\sqrt{Re}/M_0$	Free Molecule Eq. (4.35)	Rayleigh Eq. (4.37)	small time Eq. (4.32)	large time Eq. (4.33)	Approximate Eq. (4.34)
0	.683		.683		.683
.01			.683		.681
.1			.683		.658
.316			.677		.607
.5			.668		.569
1			.627		.483
1.414			.583		.427
2		.564	.532	1.923	.364
3		.376	.612	.534	.287
4		.282		.313	.236
5		.226		.232	.198
8		.141		.141	.134
10		.113		.112	.110
100		.0113		.0113	.0113
1,000		.00113		.00113	.00113
10,000		.000113		.000113	.000113
∞		0		0	0

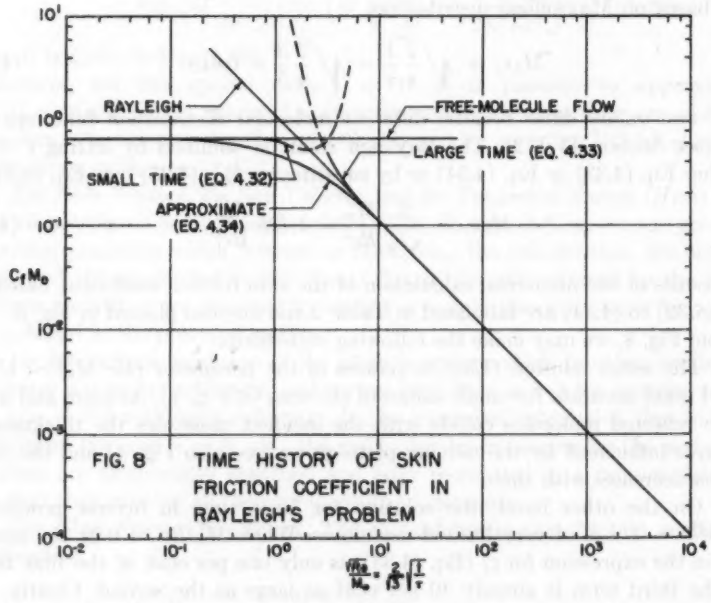


FIG. 8 TIME HISTORY OF SKIN FRICTION COEFFICIENT IN RAYLEIGH'S PROBLEM

all times. This approximate solution approaches the exact free molecule solution and the exact Rayleigh solution at the respective limits and gives the right trend of the skin friction, which is enveloped by the free molecule and the Rayleigh values. It is to be noted that the approximate solution (4.34) might slightly under-estimate the skin friction coefficient.

2. The Slip Velocity and the Tangential Energy (Heat) Flux. The slip velocity is defined as the difference between the velocity of the flat plate and the velocity of the gas adjacent to the plate, $U - u(t, 0)$. The Laplace transform of this quantity is $U(1/s) - \bar{u}(s, 0)$. From Eqs. (4.28) and (4.29), we obtain the approximation

$$U\frac{1}{s} - \bar{u}(s,0) = U\frac{1}{N^{-1}s + G^{-1}\sqrt{\frac{p_0}{\mu_0}}\sqrt{s}} \tag{4.38}$$

where

$$N = \frac{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{7\pi}{10}} + \dfrac{1}{5}}{\dfrac{1+\alpha}{1-\alpha}\sqrt{\dfrac{7\pi}{10}} + \dfrac{6}{5}}$$

and G is given in Eq. (4.18). In the denominator of Eq. (4.38) the first term is the dominant term for large s or small t, and the second is the dominant term for small s or large t. These two terms are not of the same order, so the approximation is possible for the whole s range, and its inversion would be valid for all times.

By inverting* Eq. (4.38), we have

$$U - u(t,0) = UN \exp\left(\frac{N^2}{G^2}\frac{p_0}{\mu_0}t\right)\operatorname{erfc.}\left(\frac{N}{G}\sqrt{\frac{p_0}{\mu_0}t}\right) \tag{4.39}$$

where G is given in Eq. (4.18), and N is given in Eq. (4.38).

Putting $y = 0$ and $\alpha = 0$ in Eqs. (4.15) and (4.20), $\alpha = 0$ in Eq. (4.39), and using Eq. (4.17) we obtain the following solutions for the slip velocity:

$$1 - \frac{u(t,0)}{U} = 0.627 - 0.0847\frac{Re}{M_0^2} + 0.0105\frac{Re^2}{M_0^4} + 0\left(\frac{Re^3}{M_0^6}\right) \quad t \text{ small} \tag{4.40}$$

$$1 - \frac{u(t,0)}{U} = 0.913\frac{M_0}{Re^{\frac{1}{2}}} - 2.086\frac{M_0^3}{Re^{\frac{3}{2}}} + 0.147\frac{M_0^5}{Re^{\frac{5}{2}}} + 0\left(\frac{M_0^7}{Re^{\frac{7}{2}}}\right) \quad t \text{ large} \tag{4.41}$$

$$1 - \frac{u(t,0)}{U} \cong 0.627\exp\left(0.388\frac{\sqrt{Re}}{M_0}\right)^2 \operatorname{erfc.}\left(0.388\frac{\sqrt{Re}}{M_0}\right) \quad \text{all } t \tag{4.42}$$

The results of the numerical calculations are tabulated in Table 3 and are plotted in Fig. 9.

* See Table of Transforms, Formula 43, p. 297, Reference 24.

TABLE 3

Time History of Slip Velocity $1 - u(t, 0)/U$ *in Rayleigh's Problem*

$\sqrt{Re}/M_0$	small time Eq. (4.40)	large time Eq. (4.41)	Approximate Eq. (4.42)
0	.627		.627
.01	.627		.625
.1	.626		.601
.316	.619		.549
.5	.607		.511
1	.553		.425
1.414	.500		.371
2	.456	.200	.312
3	.712	.228	.242
4		.196	.197
5		.166	.165
8		.110	.110
10		.089	.089
100		.009	.009
1,000		.0009	.0009
10,000		.000009	.000009
∞		0	0

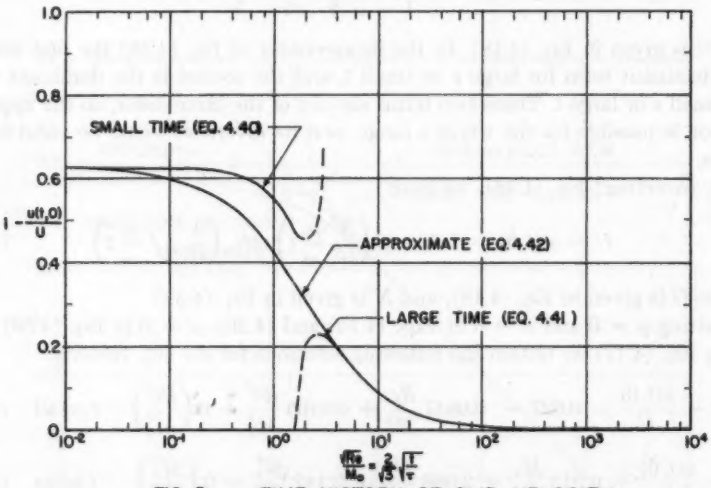


FIG. 9 TIME HISTORY OF SLIP VELOCITY IN THE RAYLEIGH PROBLEM

FIG. 9

The velocity slip at zero time is 63% rather than the conventional 50% for free molecule flow based on the Maxwellian distribution far from the plate surface. According to Grad's approximate solution of the Maxwell-Boltzmann equation, the distribution function experiences a "jump" at $t = 0^+$, and this fact accounts for the difference between the present results and those of conventional free-molecule flow. (See also SectionV.)

According to Fig. 9, the velocity slip is less than 10% when $\sqrt{Re}/M_0 > 10$, and less than 1% when $\sqrt{Re}/M_0 > 100$. For $\sqrt{Re}/M_0 > 100$, the no-slip boundary condition certainly applies. The approximate formula (4.42) is seen to be somewhat less accurate than that for the skin friction.

Similarly, the energy (heat) flux parallel to the plate may be obtained and is plotted in Fig. 10 and tabulated in Table 4.

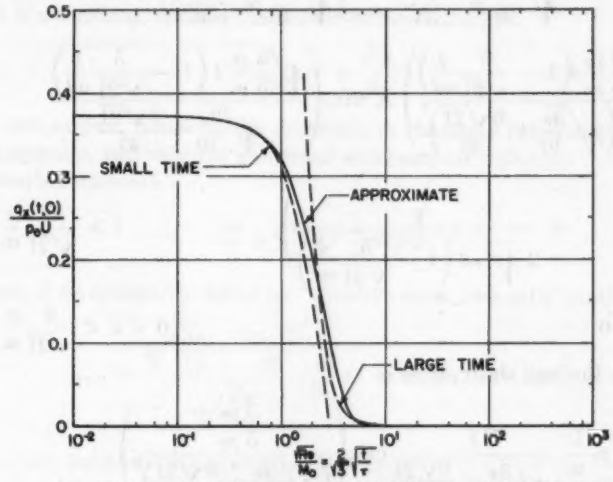


FIG. 10 TIME HISTORY OF TANGENTIAL ENERGY (HEAT) FLUX IN THE RAYLEIGH PROBLEM

FIG. 10

TABLE 4

Time History of Tangential Energy (Heat) Flux $q_x(t, 0)/p_0U$ in Rayleigh's Problem

$\sqrt{Re}/M_0$	small time	large time	Approximate
0	.373		.373
.01	.373		.373
.1	.372		.372
.316	.366		.367
.5	.357		.359
1	.311		.321
1.414	.254		.277
2	.154	.242	.205
3		.0508	.097
4		.0199	.034
5		.0098	.0089
8		.00229	.00003
10		.00116	.000000
100		.000001	.000000
1,000		.000000	.000000
10,000		.000000	.000000
∞		0	0

F. Comparison with Schaaf's Solution for Wall Shear Stress $p_{xy}(t, 0)$. The closed-form approximate solution for the shear stress is given by Eq. (4.23). For the special case of $\alpha = 0$, we have

$$p_{xy}(t,y) = p_0 \frac{U}{a_0} \frac{1}{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}} \exp\left(\frac{\sqrt{\frac{3}{5}}\sqrt{\frac{p_0}{\mu_0}}\frac{y}{\sqrt{\nu_0}}}{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}}\right)$$

$$\cdot \exp\left(\frac{\frac{3}{5}\frac{p_0}{\mu_0} t\left(1 - \frac{5}{\sqrt{21}}\frac{y}{a_0 t}\right)}{\left(\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}\right)^2}\right) \times \operatorname{erfc}.\left(\frac{\sqrt{\frac{3}{5}\frac{p_0}{\mu_0} t\left(1 - \frac{5}{\sqrt{21}}\frac{y}{a_0 t}\right)}}{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}}\right. \tag{4.43}$$

$$\left. + \frac{y}{2\sqrt{\nu_0 t\left(1 - \frac{5}{\sqrt{21}}\frac{y}{a_0 t}\right)}}\right), \qquad t > \frac{5}{\sqrt{21}}\frac{y}{a_0}$$

$$= 0 \qquad 0 < t < \frac{5}{\sqrt{21}}\frac{y}{a_0}$$

from which the wall shear stress is

$$p_{xy}(t,0) = p_0 \frac{U}{a_0} \frac{1}{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}} \exp\left(\frac{\frac{3}{5}\frac{p_0}{\mu_0} t}{\left(\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}\right)^2}\right) \tag{4.44}$$

$$\cdot \operatorname{erfc}.\left(\frac{\sqrt{\frac{3}{5}\frac{p_0}{\mu_0} t}}{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}}\right) \qquad \text{all } t$$

The corresponding quantities obtained by Schaaf[5], using the linearized Navier-Stokes equation and a slip boundary condition, are, respectively,

$$p_{xy}(t, y) = \frac{\mu_0 U}{L'} e^{y/L'} e^{(\nu_0/L'^2)t} \operatorname{erfc}.\left(\frac{\sqrt{\nu_0 t}}{L'} + \frac{y}{2\sqrt{\nu_0 t}}\right) \tag{4.45}$$

$$p_{xy}(t, 0) = \frac{\mu_0 U}{L'} e^{(\nu_0/L'^2)t} \operatorname{erfc}.\left(\frac{\sqrt{\nu_0 t}}{L'}\right) \tag{4.46}$$

where L' is the mean free path used in Reference 5. The expressions for $p_{xy}(t, 0)$ in Eqs. (4.44) and (4.46) are of the same form, but the numerical constants are different. However, the expression (4.43) and (4.45) for $p_{xy}(t, y)$ away from the surface are very different. Our solution of Grad equations contains the inherent wave characteristics in the term

$$1 - \frac{5}{\sqrt{21}}\frac{y}{a_0 t}$$

while Schaaf's result (4.45) is diffusive in nature. Of course, for sufficiently large time, $p_{xy}(t, y)$ given by Eq. (4.43) is also diffusive, as shown in Eq. (4.26). In other words, near the start of the motion the shear stress satisfies a wave equation of the form

$$\frac{\partial^2 p_{xy}}{\partial t^2} = b^2 \frac{\partial^2 p_{xy}}{\partial y^2},$$

where b is a constant, and *not* the Navier-Stokes equation

$$\frac{\partial p_{xy}}{\partial t} = \nu_0 \frac{\partial^2 p_{xy}}{\partial y^2}$$

At the plate surface, however, the difference in character between the two solutions disappears, and only the numerical constants are different.

In Grad's framework

$$\frac{\mu_0}{L} = \frac{5}{16}\sqrt{\frac{2}{\pi}}\sqrt{p_0\rho_0}$$

Therefore, if we arbitrarily define an "effective mean free path" so that

$$L_{\text{eff}} = \frac{\sqrt{\frac{3\pi}{10}} + \frac{6\sqrt{21}}{35}}{\sqrt{\frac{3\pi}{10}}} \frac{16}{5\pi} L' = 1.843L' = 1.809L,$$

and solve the linearized Navier-Stokes equations

$$p_{xy} = -\mu_0 \frac{\partial u}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial t} = \nu_0 \frac{\partial^2 u}{\partial y^2},$$

with the slip boundary condition

$$U = u(t, 0) = -L_{\text{eff}} \left(\frac{\partial u}{\partial y}\right)_{y=0}$$

we obtain an expression for $p_{xy}(t, 0)$ identical to Eq. (4.44).

At the present time it is not clear whether the result obtained by this purely formal process is coincidental and restricted to Rayleigh's problem, or is suggestive for more general flow problems.

V. Discussion and Conclusion. According to the solution of Rayleigh's problem by means of Grad's thirteen-moment approximation, a time interval corresponding to 50–100 collisions per molecule is required to transform the initially-discontinuous, step-function distribution of gas velocity into the classical Rayleigh distribution with a negligibly small "slip" at the surface.* As anticipated from qualitative physical considerations (Introduction) the flow behaves initially like a free-molecule flow, and all physical quantities are analytic func-

* This time interval is about 10 times longer than the transit time across a steady, plane shock wave of moderate strength in a monatomic gas.

tions of the parameter t/τ, or Re/M_0^2. Although the solutions for $t/\tau = 0(1)$ seem to be satisfactory in most respects, the persistence of a sharp wave-front carrying discontinuities in u, p_{xy}, and q_x is somewhat surprising. This behavior appears to be inherent in Grad's method. One might expect that some of the fastest molecules near the plate surface would always be able to travel ahead of the wave, so that the wave front would be smoothed out. (Of course steep gradients would be expected around some "average" wave front.) For this reason (and others) it is desirable to investigate Rayleigh's problem with the aid of the linearized Maxwell-Boltzmann equation, as Wang Chang and Uhlenbeck[27] have done for some simple steady flows. Naturally this study will be guided by the results of the present investigation.

In spite of this criticism of Grad's method, the present results suggest that other simple but fundamental non-steady flow problems should be explored by means of the linearized Grad equations. In addition, the character of the solutions obtained for $t/\tau = 0(1)$ encourages us to believe that Rayleigh's problem with arbitrary impulsive velocity (high Mach number) can be attacked by means of series expansions in t/τ for "small time", combined with solutions of the Navier-Stokes equations for large time.

The character of the present solutions of Rayleigh's problem for large time adds to the growing lack of confidence in the Chapman-Enskog-Burnett method of successive approximations. Unlike Wang Chang and Uhlenbeck*, we were able to obtain asymptotic series expansions in powers of mean free path (more precisely, M_0^2/Re), but just when the deviations from Rayleigh's solution become interesting the convergence of these series is so poor that they are practically useless. Grad's method avoids this difficulty, because the explicit relation between the distribution function and the stress and energy flux quantities in no way implies a series expansion of the Burnett type. By utilizing this "freedom" we are able to obtain approximate solutions of Rayleigh's problem for low Mach number that are valid for all values of the time. It remains to exploit this technique for more general flow problems.

Appendix. Laplace Transforms of v', ρ', p', T', p'_{xx}, and q'_y. From Eqs. (3.22) (3.24), (3.25), (3.26), (3.28), and (3.30) together with boundary conditions (3.32) and (3.34), we have

$$\bar{p}' = B(s)\left[\exp\left(-\sqrt{g(s)\{f_1(s)+f_2(s)\}}\,y\right) - \sqrt{\frac{f_1(s)+f_2(s)}{f_1(s)-f_2(s)}}\cdot\exp\left(-\sqrt{g(s)\{f_1(s)-f_2(s)\}}\,y\right)\right] \tag{A.1}$$

* For the problem of heat conduction between two infinite parallel flat plates, and also for the planar Couette flow, the solutions obtained by Wang Chang and Uhlenbeck are not analytic in the limit of small mean free path ($d/L \to \infty$) where "d" is spacing between plates. This behavior may be connected with the particular geometry involved. Even in the Knudsen limit ($d/L \to \infty$) they experienced convergence difficulties, for which they offer a simple physical explanation, but these difficulties did not appear in the Couette flow between two rotating cylinders.

$$\bar{p}'_{xx} = \frac{4}{5}\frac{s}{s + \frac{p_0}{\mu_0}}\bar{p}' \tag{A.2}$$

$$\bar{p}'_{xx} = -\tfrac{1}{2}\bar{p}'_{yy} \tag{A.3}$$

$$\bar{v}' = B(s)\frac{1}{\rho_0}\frac{\frac{9}{5}s + \frac{p_0}{\mu_0}}{s + \frac{p_0}{\mu_0}} s\sqrt{g(s)\{f_1(s) + f_2(s)\}} \tag{A.4}$$

$$\times [\exp(-\sqrt{g(s)\{f_1(s) + f_2(s)\}}y) - \exp(-\sqrt{g(s)\{f_1(s) - f_2(s)\}}y)]$$

$$\bar{p}' = B(s)\frac{1}{2}\frac{\rho_0}{p_0}\frac{1}{s\left(s + \frac{p_0}{\mu_0}\right)}[(f_1(s) + f_2(s))\exp(-\sqrt{g(s)\{f_1(s) + f_2(s)\}}y) \tag{A.5}$$

$$-\sqrt{f_1^2(s) - f_2^2(s)}\exp(-\sqrt{g(s)\{f_1(s) - f_2(s)\}}y)$$

$$\bar{T}' = \frac{1}{R\rho_0}\bar{p}' - \frac{T_0}{\rho_0}\bar{p}' \tag{A.6}$$

$$\bar{q}'_y = -B(s)\frac{p_0}{\rho_0}\frac{1}{\left(s + \frac{p_0}{\mu_0}\right)\left(s + \frac{2}{3}\frac{p_0}{\mu_0}\right)}\sqrt{g(s)\{f_1(s) + f_2(s)\}}$$

$$\cdot\left[\left(\frac{33}{10}s + \frac{5}{2}\frac{p_0}{\mu_0}\right)\left\{\exp(-\sqrt{g(s)\{f_1(s) + f_2(s)y}\right.\right.$$

$$\left.\left. - \exp(-\sqrt{g(s)\{f_1(s) - f_2(s)\}}y)\right\} - \frac{5}{4}\frac{1}{s}\{(f_1(s) + f_2(s)\right. \tag{A.7}$$

$$\cdot \exp(-\sqrt{g(s)\{f_1(s) + f_2(s)\}}y) - (f_1(s) - f_2(s))$$

$$\left.\times \exp(-\sqrt{g(s)\{f_1(s) - f_2(s)\}}y)\}\right]$$

where

$$f_1(s) = \frac{78}{25}s^2 + \frac{16}{5}\frac{p_0}{\mu_0}s + \frac{2}{3}\frac{p_0^2}{\mu_0^2} \tag{A.8}$$

$$f_2(s) = \sqrt{\frac{3384}{625}s^4 + \frac{1296}{125}\frac{p_0}{\mu_0}s^3 + \frac{188}{25}\frac{p_0^2}{\mu_0^2}s^2 + \frac{8}{3}\frac{p_0^3}{\mu_0^3}s^3 + \frac{4}{9}\frac{p_0^4}{\mu_0^4}} \tag{A.9}$$

$$g(s) = \frac{\frac{\rho_0}{p_0}\frac{s}{2}}{\frac{9}{5}s + \frac{p_0}{\mu_0}} \tag{A.10}$$

and

$$B(s) = \left(1 - \frac{T_w}{T_0}\right)\frac{1}{s}\left(s + \frac{p_0}{\mu_0}\right)\Bigg/\Bigg[\frac{5}{8}\frac{1+\alpha}{1-\alpha}\left(\frac{2\pi}{RT_0}\right)^{\frac{1}{2}}\frac{1}{\rho_0}\frac{1}{s\left(s + \frac{2}{3}\frac{p_0}{\mu_0}\right)}$$
$$\sqrt{g(s)\{f_1(s) + f_2(s)\}}f_2(s) + \frac{s}{5}\left(\sqrt{\frac{f_1(s) + f_2(s)}{f_1(s) - f_2(s)}} - 1\right)\Bigg]^{-1} \tag{A.11}$$

Acknowledgments. The authors wish to express their indebtedness to Dr. Henry T. Nagamatsu, who initiated and encouraged the general study of problems in the theory of rarefied gases. The authors have also benefited by discussions with Drs. Ting-Yi Li, Julian Cole, Theodore Y. T. Wu, and others.

The junior author is grateful to Dr. William N. Lacey, Dean of Graduate Studies and to Dr. Clark B. Millikan, Director of the Guggenheim Aeronautical Laboratory, whose kindness and helpfulness have made his stay at the California Institute possible.

REFERENCES

1. Lord Rayleigh, "On the Motion of Solid Bodies through Viscous Liquid", Phil. Mag., Ser. 6, Vol. 21, pp. 647–711, 1911. Scientific Papers, Vol. 6, p. 29, Cambridge University Press, 1920.
   Also: Durand, W. F., "Aerodynamic Theory", 1st Ed., Vol. 3, p. 64, Julius Springer, Berlin, 1934.
   Also: Goldstein, S., "Modern Developments in Fluid Dynamics", 1st Ed., Vol. 1, p. 137, Oxford University Press, 1938.
2. Howarth, L., "Some Aspects of Rayleigh's Problem for a Compressible Fluid", Quart. Journ. Mech. and Applied Math., Vol. 4, Pt. 2, pp. 157–164, 1951.
3. Van Dyke, M. D., "Impulsive Motion of an Infinite Plate in a Viscous Compressible Fluid", Journ. of Applied Mathematics and Physics, (ZAMP), Vol. 3, Fasc. 5, pp. 343–353, 1952.
4. Stewartson, K., "On the Motion of a Plate at High Speed in a Viscous Compressible Fluid, I. Impulsive Motion", Proceedings of the Cambridge Philosophical Society, Vol. 51, Part 1, pp. 202–219, January, 1955.
5. Schaaf, S. A., "A Note on the Flat Plate Drag Coefficient", Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-66, Feb., 1950.
6. Sherman, F. S., "Skin Friction in Supersonic Low Density Gas Flow", Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-70, July, 1950.
———, "Skin Friction in Supersonic Low Density Gas Flow", Part 2 Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-81, Mar., 1951.
———, "Skin Friction in Subsonic Low Density Gas Flow", Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-105, July, 1952.
   Also: Schaaf, S. A. and Sherman, F. S., "Skin Friction in Slip Flow", Journ. of the Aero. Sci., Vol. 21, No. 2, pp. 85–90, Feb., 1954.
7. Mirels, H., "Estimate of Slip Effect on Compressible Laminar Boundary Layer Skin Friction", NACA TN 2609, Jan., 1952.
8. Drake, R. M., Jr. and Kane, E. D., "A Summary of the Present Status of Heat Transfer in Rarefied Gases", Appendix I, Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-73, Oct., 1950.
9. Drake, R. M., Jr.; Sauer, F. M., Jr.; and Schaaf, S. A., "Forced Convection Heat Transfer from Cylinders and Spheres in a Rarefied Gas", Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-74, Nov., 1950.

10. Drake, R. M., Jr. and Maslach, G. J., "Heat Transfer from Right Circular Cones to a Rarefied Gas in Supersonic Flow", Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-91, Apr., 1952.
11. Maslen, S. H., "Second Approximation to Laminar Compressible Boundary Layer on Flat Plate in Slip Flow", NACA TN 2818, Nov., 1952.
12. Nonweiler, T., "The Laminar Boundary Layer in Slip Flow", Report No. 62, the College of Aeronautics, Cranfield, England, Nov., 1952.
13. Jaffé, G., "Zur Methodik der kinetischen Gastheorie", Annalen der Physik, Series 5, Volume VI, pp. 195–252, 1930.
14. Keller, J. B., "On the Solution of the Boltzmann Equation for Rarefied Gases", Communication on Applied Mathematics, Vol. I, No. 3, pp. 275–286, Sept., 1948.
15. Heineman, M., "Theory of Drag in Highly Rarefied Gases", Communication on Applied Mathematics, Vol. I, No. 3, pp. 259–273, Sept., 1948.
16. Grad, H., "On the Kinetic Theory of Rarefied Gases", Communications on Pure and Applied Mathematics", Vol. 4, No. 4, pp. 331–407, Dec., 1949.
Also: Chapman, S., and Cowling, T. G., "The Mathematical Theory of Non-Uniform Gases", 2nd Ed., p. 404, Cambridge University Press, 1952.
17. Hurlbut, F. C., "An Experimental Molecular Beam Investigation of the Scattering of Molecules from Surfaces", Univ. of Calif., Inst. of Eng. Res., Report No. HE-150-118, Dec., 1953.
18. Bell, S. and Schaaf, S. A., "Aerodynamic Forces on a Cylinder for the Free Molecule Flow of a Non-Uniform Gas", Journ. of the American Rocket Society, pp. 314–322, Sept.-Oct., 1953.
19. Grad, H., "The Profile of a Steady Plane Shock Wave", Communication on Pure and Applied Mathematics, Vol. 5, No. 3, pp. 257–300, Aug., 1952.
20. Rose, M., "Cylindrical Couette Flow", Physical Review, Vol. 91, No. 2, p. 469, July 15, 1953.
21. Gilbarg, D. and Paolucci, D., "The Structure of Shock Waves in the Continuum Theory of Fluids", Journ. of Rational Mechanics and Analysis, Vol. 2, No. 4, pp. 617–642, 1953.
22. Sherman, F. S., "A Low Density Wind Tunnel Study of Shock Wave Structure and Relaxation Phenomena in Gases", NACA TN 3298, July, 1955.
23. Jeans, J., "An Introduction to the Kinetic Theory of Gases", 1st Ed., p. 247, Cambridge University Press, 1952.
24. Churchill, R. V., "Modern Operational Mathematics in Engineering," 1st Ed., pp. 294–302, McGraw-Hill Book Co., Inc., New York, 1944.
25. Carslaw, H. S. and Jaeger, J. C., "Operational Methods in Applied Mathematics," 2nd Ed., pp. 271, 282, Oxford University Press, 1949.
26. Rosser, J. B., "Theory and Application of $\int_0^s e^{-x^2}\,dx$ and $\int_0^x e^{-p^2y^2}\,dy \int_0^y e^{-x^2}\,dx$. Part I. Methods of Computation", Final Report B2.1 from Allegany Ballistic Laboratory OSRD 5861, 1945. Also published by Mapleton House, Brooklyn, New York, 1948.
27. Wang Chang, C. S. and Uhlenbeck, G. E., "The Heat Transport between Two Parallel Plates as Functions of the Knudsen Number", Univ. of Mich., Eng. Res. Inst., Project M 999, Sept., 1953.
——, "The Couette Flow between Two Parallel Plates as a Function of the Knudsen Number", Univ. of Mich., Eng. Res. Inst., Project 1999, June, 1954.

Guggenheim Aeronautical Laboratory
California Institute of Technology
Pasadena, California*

(Received August 24, 1955)

* Dr. Yang is now at the Institute for Fluid Dynamics and Applied Mathematics, Univ. of Maryland.

A FREE STREAMLINE THEORY FOR TWO-DIMENSIONAL FULLY CAVITATED HYDROFOILS*

By T. Yao-tsu Wu

I. Introduction. The problem of cavity flows received attention early in the development of hydrodynamics because of its occurrence in high speed motion of solid bodies in water. Many previous works in this field were mainly concerned with the calculation of drag in a cavitating flow. The lifting problem with a cavity (or wake) arose later in the applications of water pumps, marine propellers, stalling airfoils, and hydrofoil crafts. Although several formulations of the problem of lift in cavity flows have been pointed out before (1–3), these theories have not yet been developed to yield general results in explicit form so that a unified discussion can be made.

The problems of cavitating flow with finite cavity demand an extension of the classical Helmholtz free boundary theory for which the cavity is infinite in extent. For this purpose, several self-consistent models have been introduced, all aiming to account for the cavity base pressure which is in general always less than the free stream pressure. In the Helmholtz-Kirchhoff flow these two pressures are assumed equal.

Of all these existing models, three significant ones may be mentioned here. The first representation of a finite cavity was proposed by Riabouchinsky (4) in which the finite cavity is obtained by introducing an "image" obstacle downstream of the real body. A different representation in which a reentrant jet is postulated was suggested by Prandtl, Wagner, and was later considered by Kreisel (5) and was further extended by Gilbarg and Serrin (3). Another representation of a free streamline flow with the base pressure different from the free stream pressure was proposed recently by Roshko (6). In this model the base pressure in the wake (or cavity) near the body can take any assigned value. From a certain point in the wake, which can be determined from the theory, the flow downstream is supposed to be dissipated in such a way that the pressure increases gradually from the assigned value to that of the free stream in a strip parallel to the free stream. Apparently this model was also considered independently by Eppler (7) in some generality. Other alternatives to these models have also been proposed (8), but they do not differ so basically from the above three models that they need to be mentioned here specifically. The mathematical solutions to the problem of flow past a flat plate set normal to the stream have been carried out for these three models (9, 6). All the theories are found to give essentially the same results over the practical range of the wake underpressure. That such agreement is to be expected can be indicated, without the detailed solutions for the various models, from consideration of their underlying physical significance, as will be discussed in the next section.

* This study was supported by the U. S. Navy, Office of Naval Research, under Contract N6onr-24420 (NR 062-059). Reproduction in whole or in part is permitted for any purpose of the U. S. Government.

In the present work the free streamline theory is extended and applied to the lifting problem for two-dimensional hydrofoils with a fully cavitating wake. The analysis is carried out by using the Roshko model to approximate the wake far downstream. The reason for using this model is mainly because of its mathematical simplicity as compared with the Riabouchinsky model, or the reentrant jet model. In fact, it can be verified that these different models all yield practically the same result, as in the pure drag case; the deviation from the results of one model to another is not appreciable up to second order small quantities. The mathematical considerations here, as in the classical theory, depend on the conformal mapping of the complex velocity plane into the plane of complex potential. By using a generalization of Levi-Civita's method for curved barriers in cavity flows, the flow problem for curved hydrofoils is finally reduced to a nonlinear boundary value problem for an analytic function defined in the upper half of a unit circle to which the Schwarz's principle of reflection can be applied. The problem is then solved by using the expansion of this analytic function inside the unit circle together with the boundary conditions in the physical plane. In order to avoid the difficulty in determining the separation point of the free streamline from a hydrofoil with blunt nose, the hydrofoils investigated here are those with sharp leading and trailing edges which are assumed to be the separation points. Except for this limitation, the present nonlinear theory is applicable to hydrofoils of any geometric profile, operating at any cavitation number, and for almost all angles of attack as long as the wake has a fully cavitating configuration.

As two typical examples, the problem is solved in explicit form for the circular arc and the flat plate for which the various flow quantities are expressed by simple formulas. From the final result the various effects, such as that of cavitation number, camber of the profile and the attack angle, are discussed in detail. It is also shown that the present theory is in good agreement with the experiment.

II. Remarks on Models in Free Streamline Theory. The total force (drag and lift for two-dimensional flows) exerted by the fluid on a solid body may of course be expressed as an integral of the local force, which consists of both the pressure and viscous stresses over the surface of the solid. However, it is also possible to express the total force in terms of integrals over surfaces at a distance from the body by applying the momentum theory. In the case of the real fluid flow past a bluff body, experimental observations indicate that the discontinuous surfaces in the flow, or free streamlines, are actually thin shear layers, into which the vorticity is fed from the boundary layer in front of the separation point. The shear layers in general do not continue smoothly far downstream, but roll up to form vortices, alternately on each side with a certain frequency. These vortices diffuse rapidly and are eventually dissipated in the wake. With a constant upsteam velocity, the wake flow is thus only stationary in the mean. Because of this complicated wake flow, it seems rather fruitless to apply the momentum theory which requires a detailed consideration of the free streamlines

at infinity. A more realistic way to formulate the problem is thus to obtain a solution which is accurate near the body by taking appropriate consideration of the cavity pressure. It is physically plausible that the detailed structure of the wake far downstream has indeed negligible influence upon the flow field near the body. Consequently, one may represent the dissipative wake flow by an equivalent model of potential flow, if properly chosen.

The general character of cavity flows depends on the value of the cavitation number σ which is defined by

$$\sigma = (P - p_c)/(\tfrac{1}{2}\rho U^2) \tag{2.1}$$

where P denotes the pressure of the undisturbed free stream, U is its relative velocity, ρ is the fluid density, and p_c is the pressure of the vapor or gas in the cavity. Physical cavities usually have finite length and positive cavitation number ($\sigma > 0$, or $p_c < P$). Some mathematical formulations have been suggested previously to investigate the possibility for obtaining the solution of steady cavity flows of ideal fluid satisfying the following conditions:

(i) the cavitation number is greater than zero, $\sigma > 0$, or $p_c < P$;
(ii) the cavity pressure is uniform throughout the cavity;
(iii) the pressure of the fluid is nowhere less than the cavity pressure;
(iv) the pressure is continuous across the cavity boundary.

The first two conditions are imposed because of their physical reality ($\sigma = 0$ case is physically unreal). The third condition follows from the potential flow theory that the velocity cannot have a maximum in the interior of the flow field. The last condition states that the interface between two phases of fluid cannot withstand any pressure jump. It then follows from (ii) that the free streamlines are surfaces of constant velocity; also, (iii) always implies that the cavity boundary is convex towards the interior flow. Application of Bernoulli's equation with condition (iii) shows that the velocity is maximum on the cavity boundary. It should be noted that these conditions differ only in (i) from the classical theory, for which $\sigma = 0$ (or $p_c = P$), corresponding to an infinitely long cavity. If the flow is restricted to be everywhere potential flow, then the cavity cannot have finite length for $\sigma \geqq 0$. For if the cavity were to close up at the rear end by itself, then the free streamlines from two sides of the body should meet either at a stagnation point from opposite directions or with a cusp in the downstream direction. The first conjecture contradicts (ii), while the second alternative violates (iii). Though the streamlines may reverse in direction and form a reentrant jet, as often observed, the jet cannot terminate in the physical plane. The above argument indicates the need of models of potential flow to represent the dissipative wake downstream if the formulation of the problem is to be retained within the scope of potential theory.

Now the problem of flow past a two-dimensional flat plate set normal to the stream, forming a finite cavity, shall be reviewed to exhibit the characteristics of these models. In the model of Riabouchinsky, an image plate A' (see Fig. 1a) is put downstream of the real plate A. The free streamlines run from A to A' and the flow field is assumed irrotational everywhere. The total force on the pair

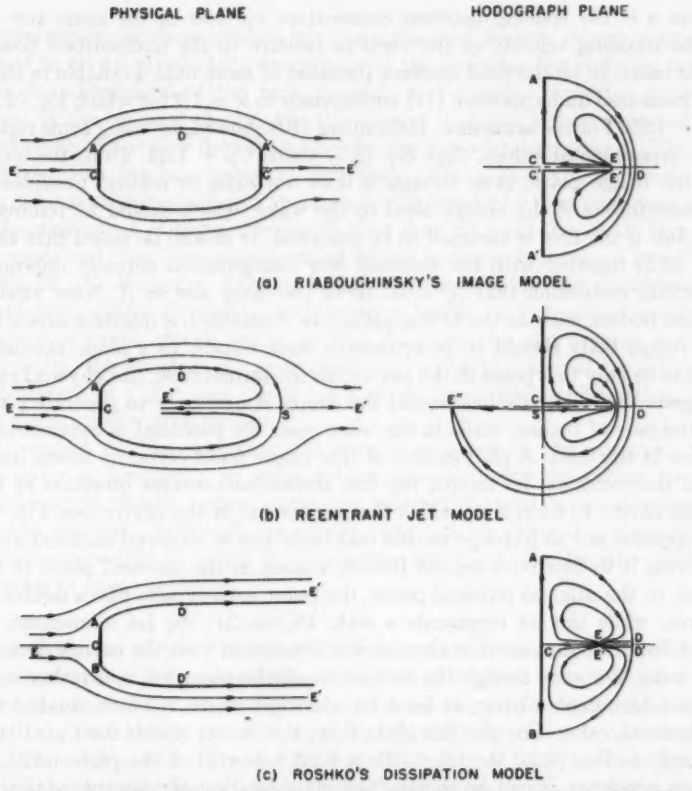


FIG. 1. Flows in the physical and hodograph plane for the various models. Note the resemblance of the streamlines in the hodograph planes near the plate AC.

of plates then vanishes, but the calculation of the drag on A alone yields the following approximate expression for the drag coefficient:

$$C_D(\sigma) \cong \{1 + \sigma + [8(\pi + 4)]^{-1}\sigma^2\}C_D(0) \cong (1 + \sigma)C_D(0) \tag{2.2}$$

where $C_D(0) = 2\pi/(4 + \pi) = 0.88$ is the classical Helmholtz solution for $\sigma = 0$. The physical significance of the image plate can perhaps be justified by the following argument. In a coordinate system where the fluid at infinity is at rest, the force holding A' does an amount of negative work, equal to $W = -DU$, where D is the drag on A and U is the speed of the traveling plates. The coefficient of work defined by $C_w = W/(\frac{1}{2}\rho U^2 l)$, where l is the plate width, then equals $-C_D$. Now if the wake is approximated by a vortex street and C_D be calculated from the vortex energy shed into the wake, the result is (see Ref. 10, p. 557)

$$C_D = \frac{a}{l}\left[1.59\ \frac{U_s}{U} - 0.63\left(\frac{U_s}{U}\right)^2\right] \tag{2.3}$$

where a is the spacing between consecutive vortices in the same row and U_s is the receding velocity of the vortices relative to the undisturbed flow. If the air is taken to be the fluid medium (because of more data available in this case), the measured underpressure (11) corresponds to $\sigma = 1.2$ for which Eq. (2.2) gives $C_D = 1.93$. For the same flow, Heisenberg (10) showed, by using some representative experimental values, that Eq. (2.3) yields $C_D = 1.82$. Thus, the work done by the image plate, even though it does not exist in reality, provides a good representation of the energy shed to the wake, which should be removed from the flow if the flow is assumed to be potential. It should be noted that the result Eq. (2.2) together with the assumed flow configuration actually imposes on A' a certain restriction that A' must be of the same size as A. Now when asymmetric bodies, such as the lifting plate, are considered, a question arises whether the image body should be in symmetry with respect to a plane normal to the stream or only to a point in the cavity. In the former case, the physical reasoning suggests that a circulation around the cavity is necessary to generate a total lift on the pair of bodies; while in the later case, the potential is continuous everywhere in the fluid. A clarification of this vague point certainly seems necessary.

In the reentrant jet model, the free streamlines reverse direction at the rear of the cavity to form a jet which flows upstream in the cavity (see Fig. 1b) and is supposed not to impinge on the real body but is removed mathematically by allowing it to flow on a second Riemann sheet in the physical plane to infinity. Thus, in the original physical plane, the point infinity acts like a doublet plus a source, while the jet represents a sink. Physically, the jet momentum carried away from the first sheet is then closely associated with the energy dissipated in the wake. For even though the jet can usually be observed, it is rather weakened by the turbulent mixing, at least its observed width is much smaller than its theoretical value. For the flat plate flow, the theory yields for $C_D(\sigma)$ the same formula as Eq. (2.2), the jet width is $0.22(1 + \sigma/4)$ of the plate width. In the lifting problems, it can be shown from consideration of momentum that the jet will turn in direction and eventually flow towards the downstream on the second sheet for small enough attack angles.

In Roshko's dissipation model, the flow past a bluff body is considered in two parts. Near the body it is described by the free streamline theory to allow a possible adjustment of the underpressure. The flow farther downstream is described by an equivalent potential flow so that its pressure increases continuously to the free stream value as it approaches infinity in a strip parallel to the free stream (see Fig. 1c). In this manner, the actual dissipative wake flow is represented; the detailed mechanism according to which the flow is dissipated is immaterial. Thus, it can be seen that this model is actually closer to the physical fact than the other models for the air flow past bluff bodies. However, this model applies equally well to cavity flows if the wake in front of the dissipation range is considered to be the front half of the cavity. By using this theory, the solution of $C_D(\sigma)$ on the flat plate differs from the previous two solutions, Eq. (2.2), only slightly by a term of $0(\sigma^2)$. Moreover, for the lifting problems, the assumption that the dissipative wake is parallel to the free stream is still physically sound and hence needs no further modification.

To summarize, all these models have one essential feature in common: they are aimed to give a satisfactory description of the flow near the body by making possible an adjustment of the base pressure and thereby removing a serious limitation in the classical theory. For this reason, it should not be expected that these solutions of the flow field shall describe the wake far downstream. The freedom of adjusting the base pressure, therefore, yields a family of solutions containing one parameter which can be expressed in terms of σ. For pure drag problems, the agreement of these theories in that they all yield the result Eq. (2.2) can be explained also from consideration of the hodograph involved (see Fig. 1 and Refs. 12, 13). However, it should be pointed out that for lifting problems the linear relation in σ for C_L and C_D such as given by Eq. (2.2) becomes invalid for moderate and small attack angles as will be shown by the present analysis. Consequently, the problem of calculating $C_L(\sigma)$ and $C_D(\sigma)$ cannot be reduced to the calculation of $C_L(0)$ and $C_D(0)$ for a given geometric configuration, as is usually done in pure drag problems.

With respect to the mathematical details involved in the analysis, these models differ in simplicity to some extent, especially for the lifting case. In Riabouchinsky's theory, some numerical integration containing elliptical integrals are required. Although the results obtained by using the reentrant jet model are expressible in terms of elementary functions, the Roshko model is still simpler in many respects.

III. Formulation of the Problem. The free streamline theory is applied here to investigate the steady, two-dimensional flow at a given attack angle α past a hydrofoil with a fully cavitating wake. The leading and trailing edges of the hydrofoil are assumed to be sharp and the flow configurations concerned are such that free streamlines separate from the hydrofoil at these sharp edges, but otherwise the wetted side of the hydrofoil may have any continuous profile. After the cavity is fully developed, the thickness of the hydrofoil has no effect on the flow and, therefore, the hydrofoil will be assumed to be of zero thickness. The Roshko model (see previous sections) will be used to approximate the wake far downstream. The flow in the physical space (or z-plane, where $z = x + iy$ with x parallel to the free stream) is shown in Fig. 2. Outside the wake the flow is assumed to be everywhere irrotational. Thus, from the complex potential

$$W(z) = \varphi(x, y) + i\psi(x, y),$$

the complex velocity w can be derived as

$$w(z) = dW/dz = u - iv = qe^{-i\theta},$$

where q, θ are the magnitude and direction of the velocity field. Here, the velocity at separation is normalized to $q = 1$, and remains at this value along the free streamlines until the latter reach points D and D' where $\theta = 0$. Downstream of these two points, the free streamlines keep parallel to the free stream on DE' and $D'E'$ along which q decreases from unity back to the free stream value U. In order that the cavitation number of the flow be σ (see Eq. (2.1) for its definition)

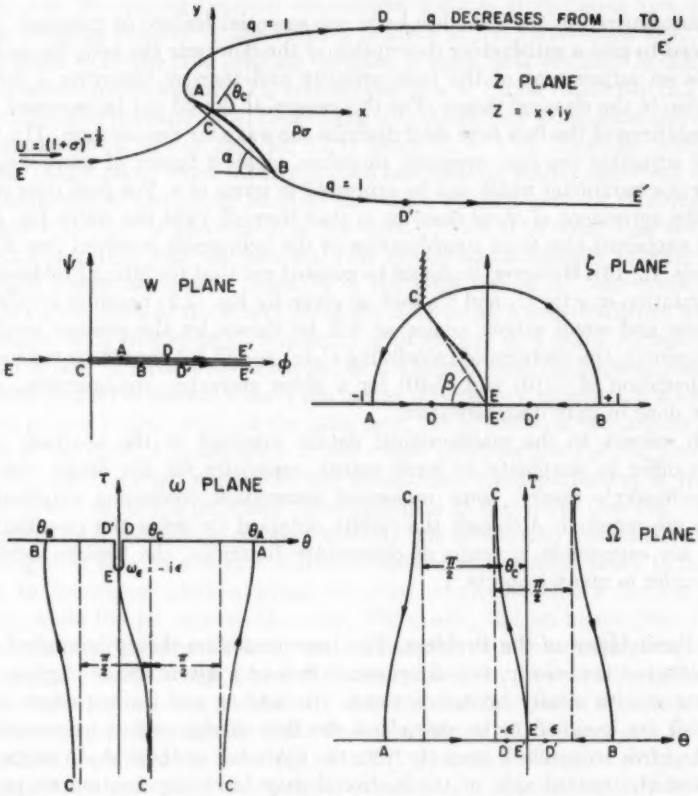


FIG. 2

with $q = 1$ on AD and BD', U takes the value $(1 + \sigma)^{-\frac{1}{2}}$, as can be shown by applying Bernoulli's equation

$$p + \tfrac{1}{2}\rho q^2 = P + \tfrac{1}{2}\rho U^2.$$

The mathematical problem can be treated by finding the conformal transformation which maps the W-plane into the w-plane. From this relation the physical plane can be deduced by an integration

$$z = \int dW/w$$

We introduce to W a transformation in ζ which is given by

$$\sqrt{W} = -b[\cos\beta + \tfrac{1}{2}(\zeta + \zeta^{-1})] \tag{3.1}$$

where

$$b = \tfrac{1}{2}(b_1 + b_2), \qquad \cos\beta = (b_2 - b_1)/(b_1 + b_2),$$

and b_1 and b_2 are two positive quantities defined by the potential at A and B

$$W_A = b_1^2, \qquad W_B = b_2^2 e^{2\pi i}.$$

This transformation maps the entire flow in the W-plane into the interior of the upper semi-circle of unit radius in the ζ-plane. The barrier ACB corresponds to the semi-circle $|\zeta| = 1, 0 \leqq \arg \zeta \leqq \pi$, and the free streamlines ADE' and $BD'E'$ map on to the two halves of the diameter along the real axis. The approaching streamline EC leaves $\zeta = 0$ perpendicular to the real axis and approaches C along the direction $\arg \zeta = \pi - \beta$.

Because of the way the nonlinear boundary conditions are specified in the z-plane (that is, with θ described on the barrier and q given on the free streamlines), we introduce the variable

$$\omega = i \log w = \theta + i \log q \equiv \theta + i\tau.$$

The flow in the ω-plane is sketched in Fig. 2. The jump in the real component θ of ω at C, where $\tau = -\infty$, follows from the fact that the two branches of the streamlines at the stagnation point C in the z-plane are orthogonal. The notch DED' in the ω-plane results from the assumption introduced in this model. At E, $\omega_E = -i\varepsilon$ where $\varepsilon = \frac{1}{2} \log (1 + \sigma)$. This notch can be removed by the transformation

$$\Omega = -\sqrt{\omega^2 + \varepsilon^2}, \qquad \varepsilon = \tfrac{1}{2} \log (1 + \sigma). \tag{3.2}$$

It should be noted that the present case ($\sigma > 0$) differs from the classical theory ($\sigma = 0$) only by the last transformation. When $\sigma = 0$, Ω and $-\omega$ become identical and the problem reduces to the classical Levi-Civita problem.

It follows from the jump condition of $\Omega(\zeta)$ at $\zeta_c = e^{i(\pi-\beta)}$ that the analytic function $\Omega(\zeta)$ has a logarithmic singularity at ζ_c and is regular elsewhere within the semi-circle. Moreover, we note that $\Omega(\zeta)$ is real for real ζ, and therefore the function $\Omega(\zeta)$ can be continued analytically over the lower half of the unit circle by applying Schwarz's principle of symmetry. More precisely, since $\Omega(0) = 0$, we may express the function $\Omega(\zeta)$ by

$$\Omega(\zeta) = i \log [(1 + \zeta e^{-i\beta})/(1 + \zeta e^{i\beta})] + \sum_{n=1}^{\infty} A_n \zeta^n. \tag{3.3}$$

The first term on the right-hand side represents the singular part of $\Omega(\zeta)$ while the series denotes the expansion of an analytic function, regular and hence convergent, in and on the unit circle. The coefficients, A_n's, are real and can be determined in principle from the geometry of the barrier ACB (see Sec. 4). Near the origin, we have

$$\Omega(\zeta) = \sum_{n=1}^{\infty} \left[(-1)^{n+1} \frac{2}{n} \sin n\beta + A_n \right] \zeta^n \equiv \sum_{n=1}^{\infty} a_n \zeta^n, \qquad (|\zeta| < 1). \tag{3.4}$$

On the barrier, where $\zeta = e^{i\eta}$, $0 \leqq \eta \leqq \pi$, we have

$$T(\eta) = \operatorname{Im} \Omega = \tfrac{1}{2} \log \frac{1 + \cos (\eta - \beta)}{1 + \cos (\eta + \beta)} + \sum_{n=1}^{\infty} A_n \sin n\eta; \tag{3.5a}$$

$$\Theta(\eta) = \mathrm{Rl}\,\Omega = \beta_0 + \beta + \sum_{n=1}^{\infty} A_n \cos n\eta; \tag{3.5b}$$

where $\beta_0 = 0$ for $0 \leqq \eta < (\pi - \beta)$ and $\beta_0 = -\pi$ for $(\pi - \beta) < \eta \leqq \pi$. Now from the definition of ω and the expression of W in terms of ζ, (3.1), the physical plane $z(\zeta)$ can be obtained by integrating

$$dz = e^{i\omega}\,dW = \frac{b^2}{2}\exp\{-i(\Omega^2 - \varepsilon^2)^{\frac{1}{2}}\}\left(\zeta + \frac{1}{\zeta} + 2\cos\beta\right)\left(\zeta - \frac{1}{\zeta}\right)\frac{d\zeta}{\zeta} \tag{3.6}$$

which, in particular, reduces on the wetted wall, $\zeta = e^{i\eta}$, to

$$z = 2b^2 \int_{\eta}^{\pi-\beta} e^{-i(\Omega^2-\varepsilon^2)^{\frac{1}{2}}}(\cos\eta + \cos\beta)\sin\eta\,d\eta. \tag{3.7}$$

The arc length of the barrier along CA or CB can be determined from

$$s = \int_{\eta}^{\pi-\beta} \frac{|dz|}{d\eta}\,d\eta = 2b^2 \int_{\eta}^{\pi-\beta} e^{-\tau(\eta)}(\cos\eta + \cos\beta)\sin\eta\,d\eta; \tag{3.8}$$

and the total length of the wetted wall is then

$$S = 2b^2 \int_0^{\pi} e^{-\tau(\eta)}\,|\cos\eta + \cos\beta|\sin\eta\,d\eta \tag{3.9}$$

where $\tau = \mathrm{Im}\,\omega = -\mathrm{Im}\,(\Omega^2 - \varepsilon^2)^{\frac{1}{2}}$. The radius of curvature of the barrier is given by

$$R = \frac{ds}{d\theta} = 2b^2 e^{-\tau(\eta)}\,|\cos\eta + \cos\beta|\sin\eta\,\frac{d\eta}{d\theta}. \tag{3.10}$$

If X denotes the drag, Y, the lift, then it can be shown that the force is given by (e.g. Ref. 14, p. 305)

$$\begin{aligned} X + iY &= -\frac{i\rho}{2}\oint e^{i\omega}\frac{dW}{d\zeta}\,d\zeta \\ &= -\frac{i\rho b^2}{4}\oint \exp\{-i(\Omega^2 - \varepsilon^2)^{\frac{1}{2}}\}\left(\zeta + \frac{1}{\zeta} + 2\cos\beta\right)\left(\zeta - \frac{1}{\zeta}\right)\frac{d\zeta}{\zeta} \end{aligned} \tag{3.11}$$

where the contour for the integral is $|\zeta| = 1$. Likewise the moment M of the force about the stagnation point C, positive in the sense of nose-down, is found to be (14)

$$M = \frac{\rho}{2}\,\mathrm{Rl}\int_{(BCA)} [e^{-i\omega(\zeta)} - e^{-i\omega(\bar{\zeta})}]z\frac{dW}{d\zeta}\,d\zeta; \tag{3.12}$$

the integration is taken around the semi-circle BCA in the ζ-plane. The above integral cannot be reduced to one on a closed contour, and, therefore, must be evaluated separately.

The above formulas are merely formal representations of various physical quantities. The magnification factor b for the physical plane may be eliminated from Eqs. (3.8)–(3.10). If the boundary value $R(s)$ is applied to these two resulting equations, an identity containing all the coefficients A_n in the parameter

of η can be obtained. From this identity the coefficients A_n can, at least in principle, be determined. The factor b can then be expressed by (3.9) in terms of the wall length S, and in turn, X, Y, M can be calculated from Eqs. (3.11), (3.12). In practice, however, to determine the coefficients A_n from the identity would necessitate some expansion of the functions of η involved in the identity into series. This procedure leads to infinitely many transcendental equations in infinitely many unknowns. In order to obtain the general result in explicit form with sufficient accuracy, the analysis is carried out by replacing the expansion in Eq. (3.3) by the three leading terms only, that is,

$$\Omega(\zeta) = i \log [(1 + \zeta e^{-i\beta})/(1 + \zeta e^{i\beta})] + A_1\zeta + A_2\zeta^2 + A_3\zeta^3. \tag{3.13}$$

If two particular points are suitably chosen on the arc ACB at which the boundary values of R and θ are applied, we get four transcendental equations from which the four unknowns, β, A_1, A_2, and A_3 can be determined, expressible in terms of the cavitation number σ, the attack angle α, and the profile shape. The rest of the physical quantities can then be calculated as shown below.

a. *Lift and Drag.* In the integral representing the force, given by Eq. (3.11), the integrand has inside the contour $|\zeta| = 1$ only one pole at $\zeta = 0$, but has, in addition, two branch points at $\Omega = \pm\varepsilon$. Besides, the function $\Omega(\zeta)$ in the integrand has on the contour a logarithmic singularity at C and its conjugate point in the ζ-plane. Thus, if a branch cut is introduced from D to D' along the real axis in both the ζ- and Ω-planes and two other branch cuts from C to its conjugate point outside of the unit circle in ζ-plane (see Fig. 2), the integrand is then one valued inside and on the contour in the cut plane. In practical applications, the values of the cavitation number σ usually fall in the range $0 < \sigma < 1$, corresponding to $0 < \varepsilon^2 < 0.123$, which can therefore be considered small. Moreover, the modulus $|\Omega|$ is always greater than ε on the contour. Therefore, the exponential function in Eq. (3.11) can be expanded in terms of the small quantity ε^2,

$$\exp\{-i(\Omega^2 - \varepsilon^2)^{\frac{1}{2}}\} = \exp\left\{-i\Omega + \frac{i\varepsilon^2}{2\Omega} + 0(\varepsilon^4)\right\} = e^{-i\Omega}\left(1 + \frac{i\varepsilon^2}{2\Omega} + 0(\varepsilon^4)\right). \tag{3.14}$$

It can be verified that the term of $0(\varepsilon^4)$ has indeed negligible value and thus may be omitted. Upon substitution of Eqs. (3.13), (3.14) into (3.11), the resulting integral is of the form to which the theorem of residues can readily be applied. From the residue of the integrand at $\zeta = 0$, one obtains the following result:

$$X = \pi\rho b^2\left(1 + \frac{\varepsilon^2}{6}\right)(\sin\beta + A_1/2)^2, \tag{3.15}$$

$$\begin{aligned} Y = \pi\rho b^2(1 + \varepsilon^2/4)\Big\{&(\sin\beta\cos\beta + A_1\cos\beta + A_2/2) \\ &+ \frac{\varepsilon^2}{4}\cos\beta\left(\sin\beta + \frac{A_1}{2}\right)^{-4}\left[\left(\sin\beta + \frac{A_1}{2}\right)^3\right. \\ &\left.\left. + \sin\beta\left(\sin\beta + \frac{A_1}{4}\right)\left(\frac{A_1}{2} + \frac{5}{4}A_2 + A_3\right)\right]\right\}. \end{aligned} \tag{3.16}$$

In Eq. (3.16) the quadratic terms in A_2 and A_3 are omitted because their contribution is negligible. It should be remarked here that Eq. (3.16) is valid when the denominator, $(\sin\beta + A_1/2)$, of the last term is greater than zero, implying further that the drag X is always positive definite. For, if this quantity vanishes, then the expansion of $\Omega(\zeta)$ near $\zeta = 0$ (see Eq. (3.4)) starts with ζ^2 and hence the mapping fails to be conformal at $\zeta = 0$, indicating that the flow configuration is basically different from what is being considered. If this quantity becomes negative, then the function $\Omega(\zeta)$ changes its branch, implying physically that the cavity shifts sides on the barrier so that different boundary conditons should be used. With the modified boundary conditions, the present formulation, however, would remain valid.

The factor b^2 in the above equation can be expressed in terms of the arc length S of the barrier. To evaluate the integral representing S in (3.9), we first expand $\exp(-\tau)$ in a way similar to that in (3.14),

$$e^{-\tau(\eta)} = \exp\{\mathrm{Im}\,(\Omega^2 - \varepsilon^2)^{\frac{1}{2}}\} = e^{T(\eta)}\left[1 + \frac{\varepsilon^2}{2}\frac{T}{\Theta^2 + T^2} + 0(\varepsilon^4)\right] \quad (3.17)$$

where $\Omega(e^{i\eta}) = \Theta(\eta) + iT(\eta)$, and Θ, T are given in Eq. (3.5) with $A_4 = A_5 = \cdots = 0$. Substituting Eq. (3.17) into (3.9), we obtain

$$S = 2b^2\int_0^\pi \left(1 + \frac{\varepsilon^2}{2}\frac{T}{\Theta^2 + T^2}\right)\left[1 + \sum_{n=1}^{3} A_n \sin n\eta\right][1 + \cos(\eta - \beta)]\sin\eta\, d\eta$$

in which the higher order terms of A_n are neglected. The function $T/(\Theta^2 + T^2)$ is positive and, as can be shown, is bounded above by a constant of order unity, which depends on the value of Θ at B but is independent of ε. Hence the contribution from the term of $0(\varepsilon^2)$ is very small relative to the first order term. Carrying out the integration, we finally obtain

$$Sb^{-2} \equiv J = 4 + \pi\sin\beta + A_1\left(\pi + \frac{8}{3}\sin\beta\right) + \frac{\pi}{2}A_2\cos\beta - \frac{8}{15}A_3\sin\beta. \quad (3.18)$$

For a flat plate set at an angle α, moving at the idealized limiting condition $\sigma = 0$ ($\varepsilon = 0$), the coefficients A's all vanish, as required by the condition of conformal mapping, and $\beta = \alpha$ (see Eq. (3.5b)), then Eqs. (3.15), (3.16) and (3.18) reduce to the classical result for an oblique lamina (see Ref. 15, p. 102).

b. *Moment of Force; Stagnation Point; Center of Pressure.* Applying a similar approximation, as described in Eq. (3.17), to Eqs. (3.12) and (3.7), and further neglecting A_3, we obtain

$$M = 2b^2\rho\,\mathrm{Rl}\int_0^\pi \exp\{i(\beta + A_1\cos\eta + A_2\cos 2\eta)\}$$
$$\cdot\,[\sin\beta + (A_1 + 2A_2\cos\eta)(1 + \cos\beta\cos\eta)](\sin^2\eta)z(\eta)\, d\eta$$

where

$$z(\eta) = 2b^2e^{-i\beta}\int_\eta^{\pi-\beta}[1 - iA_1e^{i\eta} - iA_2e^{2i\eta}][1 + \cos(\eta - \beta)]\sin\eta\, d\eta.$$

We then substitute the integral for $z(\eta)$ into the former expression, decompose accordingly the interval into two parts: $0 \leqq \eta \leqq \pi - \beta$ and $\pi - \beta \leqq \eta \leqq \pi$, and then interchange the order of integration so that the limits of integration further simplify the involved manipulation. After some tedious integrations, which are otherwise straightforward, we finally obtain

$$\begin{aligned} M/(2\pi\rho b^4) \equiv K = C_1\Big\{\cos\beta\Big[\frac{5}{8} + \sin^2\beta + A_1\Big(\frac{7}{6}\sin\beta + \frac{2}{3}\sin^3\beta - \frac{32}{45\pi}\Big)\Big] \\ + \frac{C_1}{2}\Big(\frac{\pi}{2} - \beta\Big) - A_2\Big(\frac{3}{16}\sin\beta + \frac{1}{12}\sin^3\beta - \sin^5\beta + \frac{64}{45\pi}\Big)\Big\} \\ + C_2\Big(1 + A_1 + \frac{A_2}{4}\cos\beta\Big)\Big(1 + \frac{128}{45\pi}\sin\beta\Big) \\ + \frac{A_2}{16}\cos\beta\Big(1 + \frac{5}{6}A_1\Big), \end{aligned} \tag{3.19}$$

where b^2 is given by Eq. (3.18) and

$$C_1 = \sin\beta + A_1 + \frac{A_2}{2}\cos\beta; \qquad C_2 = \tfrac{1}{4}(A_1\cos\beta + 2A_2).$$

If a profile symmetrical with respect to the central chord is set with its chord normal to the free stream, then $\beta = \pi/2$ due to symmetry (see Fig. 2). Furthermore, A_2 should vanish because in this case $\Omega(\zeta)$ is necessarily an odd function of ζ (see Eq. (3.13)). It then follows from Eq. (3.19) that the moment M about the stagnation point, which is now at the central chord, vanishes as it should be expected.

Because of the indefinite location of the stagnation point, the moment of force is usually calculated or measured referring to a fixed point. Hence, we derive here the moment M_0 about the leading edge. This further requires the calculation of the position of stagnation point. Denote the distance along the wall from the stagnation point to the leading edge by S_0; then S_0 can be calculated from Eq. (3.8) by letting the lower limit $\eta = \pi$. Proceeding in a manner similar to that described for the equation previous to (3.18), we obtain

$$S_0 \cong 2b^2\int_0^\beta (1 + A_1\sin\theta - A_2\sin 2\theta + A_3\sin 3\theta)[1 - \cos(\theta + \beta)]\sin\theta\, d\theta$$

which finally leads to

$$\begin{aligned} S_0 b^{-2} \equiv J_0 = 2(1 - \cos\beta) + \sin\beta(\beta - \sin 2\beta) \\ + A_1(\beta + \tfrac{4}{3}\sin\beta - \tfrac{2}{3}\sin 2\beta + \tfrac{1}{4}\sin 4\beta) \\ + A_2(\tfrac{1}{2}\beta\cos\beta - \tfrac{1}{2}\sin\beta + \tfrac{1}{8}\sin^3\beta - 2\sin^5\beta) \\ + \tfrac{8}{15}A_3\sin\beta[\cos\beta(2 + \sin^2\beta + 24\sin^4\beta) - 2]. \end{aligned} \tag{3.20}$$

The factor b^2 can be eliminated by using Eq. (3.18) to give

$$\mu \equiv S_0/S = J_0/J. \tag{3.21}$$

Again, for a symmetric profile normal to the free stream, $\beta = \pi/2$, and $A_2 = 0$, it can be verified from Eqs. (3.18) and (3.20) that $\mu = \frac{1}{2}$. For small values of β (corresponding to small α, as can be seen from Eq. (3.5b), Eq. (3.21) reduces to

$$\mu = \tfrac{17}{16}\beta^4 + 0(\beta^5) \tag{3.22}$$

which shows that the stagnation point is extremely close to the leading edge for small angles of attack.

Knowing the position of the stagnation point, the moment M_0 about the leading edge then becomes, for profiles of small camber,

$$M_0 = M + (Y \cos\alpha + X \sin\alpha)S_0 \tag{3.23}$$

Having obtained M_0, we can determine the location of the center of pressure, measured from the leading edge, to be approximately at

$$S_1 = M_0/(Y \cos\alpha + X \sin\alpha), \tag{3.24}$$

or, in percentage of the chord,

$$\nu = S_1/S = \mu + (M/S)(Y \cos\alpha + X \sin\alpha)^{-1}. \tag{3.25}$$

Now, consider again a symmetric profile normal to the free stream, $\alpha = \beta = \pi/2$. We have shown before that $\mu = \frac{1}{2}$ and $M = 0$, hence $\nu = \frac{1}{2}$. For small values of α (β is also small as previously explained), Eq. (3.22) states that μ is very small; Eq. (3.16) reduces to $Y \cong \pi\rho b^2(\beta + A_1 + A_2/2)$; Eq. (3.18) gives $S \cong 4b^2$; and $M \cong 2\pi b^4\rho[\frac{5}{8}(\beta + A_1 + A_2/2) + \frac{1}{4}(A_1 + 2A_2)]$ from Eq. (3.19). Thus, for small α and σ, we have approximately

$$\nu \cong \tfrac{5}{16} + \tfrac{1}{8}(A_1 + 2A_2)/(\beta + A_1 + A_2/2) + 0(\beta). \tag{3.26}$$

Hence the center of pressure ranges from $\frac{5}{16}$ to $\frac{1}{2}$ of the chord.

c. *Some Basic Features of the Free Streamlines.* In this section the shape of the free streamlines AD, BD' and the location of the points D and D' will now be calculated. First, we show that the assumed curvature form of the free streamlines, namely, concave towards the cavity, imposes certain conditions on the coefficients A_1, A_2 and A_3. Denote the distance along the streamline $\psi = 0$ by s, positive when away from the point C. On AD, $\zeta = -\xi$ with $1 \geqq \xi \geqq \xi_1$ where $-\xi_1$ is the value of ζ at D; while on BD', $\zeta = \xi$ with $1 \geqq \xi \geqq \xi_2$ where $\xi_2 = \zeta_{D'}$. Then the radius of curvature $R = -ds/d\theta$ on AD and $R = ds/d\theta$ on BD' should both be positive. We shall first consider R on AD. From the definition of $\Omega = \Theta + iT$ and $z(\zeta)$ (see Eqs. (3.2), (3.6)), ω and Ω are both real on AD, and hence

$$R = -\frac{ds}{d\theta} = -\frac{|dz|}{d\xi}\frac{d\xi}{d\Theta}\frac{d\Theta}{d\theta} = -\frac{\theta}{\Theta}\frac{|dz|}{d\xi}\Big/\frac{d\Theta}{d\xi}$$

$$= \frac{b^2}{2}\left(-\frac{\theta}{\Theta}\right)\xi^{-3}\frac{(1-\xi^2)(1-2\xi\cos\beta+\xi^2)^2}{2\sin\beta + (1-2\xi\cos\beta+\xi^2)(A_1 - 2A_2\xi + 3A_3\xi^2)}.$$

From this equation it follows that the condition $R(\xi) \geqq 0$ for $\xi \leqq 1$ requires

$$2\sin\beta + (1 - 2\xi\cos\beta + \xi^2)(A_1 - 2A_2\xi + 3A_3\xi^2) > 0. \tag{3.27}$$

Consequently $R(\xi)$ increases from $R(1) = 0$ as ξ decreases from 1. A similar requirement for positive R on BD' is, for $\xi_2 \leqq \xi \leqq 1$,

$$2 \sin \beta + (1 + 2\xi \cos \beta + \xi^2)(A_1 + 2A_2\xi + 3A_3\xi^2) > 0. \tag{3.28}$$

It will be only stated here that these two subsidiary conditions are in general satisfied owing to the facts that the coefficient of the first term in Eq. (3.4) is positive and that A_n's decrease rapidly with respect to A_1. However, no further elaboration will be made here on these points.

A parametric representation $x(\xi)$, $y(\xi)$ of the free streamline AD can be obtained by integrating Eq. (3.6) along $\zeta = -\xi$ from $\xi = 1$ to $\xi \geqq \xi_1$. On AD, $\tau = \operatorname{Im} \omega = 0$, and hence, referring to the point A, we have

$$x - x_A = -\frac{S}{2J} \int_1^\xi \xi^{-3}(1 - 2\xi \cos \beta + \xi^3)(1 - \xi^2) \cos \theta \, d\xi, \tag{3.29}$$

$$y - y_A = -\frac{S}{2J} \int_1^\xi \xi^{-3}(1 - 2\xi \cos \beta + \xi^3)(1 - \xi^2) \sin \theta \, d\xi. \tag{3.30}$$

In the above equations $\cos \theta$ and $\sin \theta$ are complicated functions of ξ,

$$\theta = -(\Theta^2 - \varepsilon^2)^{\frac{1}{2}}, \tag{3.31}$$

$$\Theta = -2 \tan^{-1} \frac{\xi \sin \beta}{1 - \xi \cos \beta} - A_1\xi + A_2\xi^2 - A_3\xi^3. \tag{3.32}$$

To simplify the calculation, we approximate Eq. (3.31) by $\theta = -(\Theta + \varepsilon)$ on AD, which should be good for θ both small and large. Then we obtain the following approximate formulas

$$\cos \theta = (1 - 2\xi \cos \beta + \xi^2 \cos 2\beta)/(1 - 2\xi \cos \beta + \xi^2),$$

$$\sin \theta = \frac{2\xi(1 - \xi \cos \beta) \sin \beta + (A_1\xi - A_2\xi^2 - \varepsilon)(1 - 2\xi \cos \beta + \xi^2 \cos 2\beta)}{1 - 2\xi \cos \beta + \xi^2}.$$

Substituting these values into Eqs. (3.29), (3.30) and carrying out the integrations, we finally obtain

$$\begin{aligned} 2J(x - x_A)/S \cong\ & \tfrac{1}{2}(1 - \xi^2)(\xi^{-2} - \cos 2\beta) \\ & - 2 \cos \beta(1 - \xi)^2\xi^{-1} + 2 \sin^2 \beta \log \xi; \end{aligned} \tag{3.33}$$

$$\begin{aligned} 2J(y - y_A)/S \cong\ & (2 \sin \beta + A_1 + 2\varepsilon \cos \beta)(1 - \xi)^2/\xi \\ & - \varepsilon\left(\frac{1 - \xi^2}{2\xi^2} + \log \xi\right) \\ & + (\sin 2\beta + 2A_1 \cos \beta + A_2 + \varepsilon \cos 2\beta) \\ & \quad [\log \xi + \tfrac{1}{2}(1 - \xi^2)] \\ & + \tfrac{1}{4}(A_1 \cos 2\beta + 2A_2 \cos \beta)(1 - \xi)^2(2 + \xi) \\ & - \tfrac{1}{4}A_2(1 - \xi^2)^2 \cos 2\beta \end{aligned} \tag{3.34}$$

To find the location of D, (x_D, y_D), we first determine the value ξ_1 corresponding to $\Omega = -\varepsilon$. Then, by letting $\Omega = -\varepsilon$ and $\zeta = -\xi_1$ in Eq. (3.4), we obtain

$$\xi_1 = \frac{\varepsilon}{a_1}\left[1 + \frac{a_2\varepsilon}{a_1^2} + 0(\varepsilon^2)\right].$$

Using this value of ξ_1, one can easily derive from Eqs. (3.33), (3.34) that

$$x_D \cong \frac{S}{J}\left\{\frac{a_1^2}{4\varepsilon^2} + 0\left(\frac{1}{\varepsilon}\right)\right\} = \frac{S}{J}\left(\frac{2\sin\beta + A_1}{\sigma}\right)^2 [1 + 0(\sigma)], \tag{3.35}$$

$$y_D \cong \frac{S}{J}\left\{\frac{a_1^2}{2\varepsilon} + 0(\log\varepsilon)\right\} = \frac{S}{J}\frac{(2\sin\beta + A_1)^2}{\sigma}[1 + 0(\sigma\log\sigma)]. \tag{3.36}$$

The above value of x_D and y_D may be regarded respectively as the half-length and half-width of the cavity. Thus we see that for σ small the cavity length is proportional to σ^{-2}, while the cavity width, proportional to σ^{-1}. It may also be seen that the free streamline near D lies close to a parabola

$$y^2/x \cong a_1^2 S/J = b^2(2\sin\beta + A_1)^2 \equiv C, \quad \text{say.}$$

Then a comparison with Eq. (3.15) shows that the drag X can be expressed by

$$X \cong \pi\rho b^2(\sin\beta + A_1/2)^2 = \pi\rho C/4$$

which is the general formula of Levi-Civita (16).

Now, in order to determine these unknown coefficients A_1, A_2, A_3 and β and thus to exhibit explicitly the effects of cavitation number σ, the attack angle α and the profile geometry, two specific examples, the circular arc and the flat plate, will be worked out below.

IV. Cavitating Hydrofoils with Circular Arc and Flat Plate Profile. Let us first consider the hydrofoil having the circular arc profile, with radius R and arc angle 2γ (see Fig. 3). The flat plate is just a special case with $\gamma \to 0$ ($R \to \infty$).

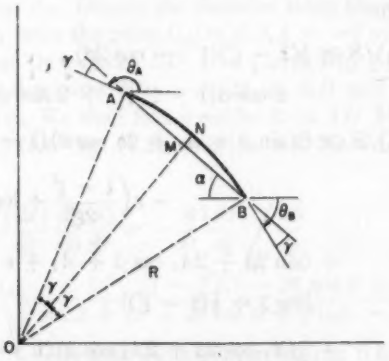


FIG. 3. The circular arc profile

We shall here restrict ourselves to small values of γ, say, $\gamma < \pi/4$. The arc length S and the length l of the chord AB are then

$$S = 2\gamma R; \qquad l = 2R \sin \gamma = 2\gamma R(1 - \gamma^2/6). \tag{4.1}$$

Now we choose the end points A and B to which the following boundary conditions are applied:

$$\begin{aligned} &\text{(i) } \theta_A = \pi - \alpha + \gamma, \qquad &&\text{(ii) } \theta_B = -\alpha - \gamma, \\ &\text{(iii), (iv) radius of curvature at } A \text{ and } B = R = S/2\gamma. \end{aligned} \tag{4.2}$$

These four conditions enable us to determine A_1, A_2, A_3 and β which in turn can be used to check the radius of curvature at other points on the boundary. Applying (i) and (ii) to the definition of θ (see Eqs. (3.2), (3.5)), we obtain

$$\Theta_B = (\theta_B^2 + \varepsilon^2)^{\frac{1}{2}} = [(\alpha + \gamma)^2 + \varepsilon^2]^{\frac{1}{2}} = \beta + A_1 + A_2 + A_3,$$

$$\Theta_A = -(\theta_A^2 + \varepsilon^2)^{\frac{1}{2}} = -\pi + \alpha - \gamma - \tfrac{1}{2}\varepsilon^2/\theta_A = -\pi + \beta - A_1 + A_2 - A_3.$$

In the second equation the value with $\frac{1}{2}$ power is expanded for θ_A is always much greater than ε, but in the first equation no expansion is made because the ratio ε/θ_B may not be small. Adding and subtracting these two equations, we have

$$A_1 + A_3 = \gamma + \tfrac{1}{2}\{[(\alpha + \gamma)^2 + \varepsilon^2]^{\frac{1}{2}} - (\alpha + \gamma)\} + \frac{\varepsilon^2}{4}(\pi - \alpha + \gamma)^{-1} \tag{4.3}$$

$$\beta + A_2 = \alpha + \tfrac{1}{2}\{[(\alpha + \gamma)^2 + \varepsilon^2]^{\frac{1}{2}} - (\alpha + \gamma)\} - \frac{\varepsilon^2}{4}(\pi - \alpha + \gamma)^{-1}. \tag{4.4}$$

In applying the conditions (iii); (iv), we first note that at B, $\tau = 0$ and

$$\frac{d\theta}{d\eta} = \frac{(\theta_B^2 + \varepsilon^2)^{\frac{1}{2}}}{\theta_B}\frac{d\Theta}{d\eta} \to -\frac{(\theta_B^2 + \varepsilon^2)^{\frac{1}{2}}}{\theta_B}(A_1 + 4A_2 + 9A_3)\sin\eta, \quad \text{as } \eta \to 0.$$

Hence, from Eq. (3.10)

$$-\frac{\theta_B(1 + \cos\beta)}{(\theta_B^2 + \varepsilon^2)^{\frac{1}{2}}}(A_1 + 4A_2 + 9A_3)^{-1} = \frac{R}{2b^2} = \frac{RJ}{2S} = \frac{J}{4\gamma};$$

or,

$$A_1 + 4A_2 + 9A_3 = (4\gamma/J)(\alpha + \gamma)[(\alpha + \gamma)^2 + \varepsilon^2]^{-\frac{1}{2}}(1 + \cos\beta). \tag{4.5}$$

Similarly, application of the boundary condition (iv) at A yields

$$A_1 - 4A_2 + 9A_3 = (4\gamma/J)(1 - \cos\beta)[1 + 0(\varepsilon^3)]. \tag{4.6}$$

In Eqs. (4.5) and (4.6), J, as defined by Eq. (3.18), also contains A_1, A_2, A_3 and β. Thus Eqs. (4.3)–(4.6) together with (3.18) represent five transcendental equations for five unknowns A_1, A_2, A_3, β and J. The solution, however, can easily be approximated with good accuracy by using an iteration procedure. For the present purpose, the following approximation is sufficient

$$\beta = \alpha + \tfrac{1}{2}\delta - \frac{\varepsilon^2}{4(\pi - \alpha + \gamma)} - \frac{\gamma}{4 + \pi \sin\alpha}\left\{\frac{\alpha + \gamma}{[(\alpha+\gamma)^2 + \varepsilon^2]^{\frac{1}{2}}}\cos^2\frac{\alpha}{2} - \sin^2\frac{\alpha}{2}\right\} \tag{4.7a}$$

$$\delta = [(\alpha + \gamma)^2 + \varepsilon^2]^{\frac{1}{2}} - (\alpha + \gamma); \tag{4.7b}$$

$$A_1 = \gamma\left\{1 + \frac{\pi \sin\alpha}{8(4 + \pi \sin\alpha)}\right\} + \frac{9}{16}\delta\left\{1 + \frac{\gamma(1 + \cos\alpha)}{9[(\alpha+\gamma)^2 + \varepsilon^2]^{\frac{1}{2}}}\right\} + \frac{9\varepsilon^2}{32(\pi - \alpha + \gamma)}; \tag{4.8}$$

$$A_2 = \frac{\gamma}{4 + \pi(A_1 + \sin\beta)}\left\{\frac{\alpha + \gamma}{[(\alpha+\gamma)^2 + \varepsilon^2]^{\frac{1}{2}}}\cos^2\frac{\beta}{2} - \sin^2\frac{\beta}{2}\right\}; \tag{4.9}$$

$$A_3 = \frac{1}{9}(\gamma - A_1) - \frac{\gamma}{9(4 + \pi \sin\alpha)}\left\{\pi \sin\beta + \frac{2\delta(1 + \cos\alpha)}{[(\alpha+\gamma)^2 + \varepsilon^2]^{\frac{1}{2}}}\right\}; \tag{4.10}$$

whereas J is still given by (3.18). The above result shows that $(\beta - \alpha)$, A_1, A_2 and A_3 are all of $0(\gamma, \varepsilon^2)$ for all α; and in particular, $(\beta - \alpha)$, A_2 and A_3 reduce to $0(\gamma\varepsilon^2, \varepsilon^4)$ for α close to $\pi/2$. Moreover, the fact that A_3 is much smaller than A_1 indicates a good accuracy of the expansion given by Eq. (3.13). If the above quantities are used to check the curvature and slope of the solid boundary at some other points, say at $\zeta = e^{i\pi/3}$, one can easily find that the agreement is within a factor at most of $0(\gamma, \varepsilon^2)$. It should also be remarked here that if more terms were taken in the expansion (3.13), then the first three coefficients A_1, A_2, A_3 would differ slightly from the above value (4.8–4.10). However, it can be verified that the "improvement" of the solution by taking terms more than three is actually unappreciable.

Now we define the conventional drag coefficient, lift coefficient and moment coefficient (about the leading edge) as follows:

$$X = \tfrac{1}{2}\rho U^2 l C_D; \qquad Y = \tfrac{1}{2}\rho U^2 l C_L; \qquad M_0 = \tfrac{1}{2}\rho U^2 l^2 C_{M_0}. \tag{4.11}$$

Then, for γ not too large ($<\pi/4$), we compile some of the previous results and note that $U = (1 + \sigma)^{-1}$, we finally obtain:

$$C_D = \frac{2\pi}{J}\left(1 + \sigma + \frac{\varepsilon^2}{6} + \frac{\gamma^2}{6}\right)\left[\sin\beta + \frac{A_1}{2}\right]^2; \tag{4.12}$$

$$C_L = \frac{2\pi}{J}\left(1 + \sigma + \frac{\varepsilon^2}{4} + \frac{\gamma^2}{6}\right)\left\{\left[\sin\beta\cos\beta + A_1\cos\beta + \frac{A_2}{2}\right] + \frac{\varepsilon^2}{4}\frac{\cos\beta}{(\sin\beta + A_1/2)^4}\left[\left(\sin\beta + \frac{A_1}{2}\right)^3 + \sin\beta\left(\sin\beta + \frac{A_1}{4}\right)\left(\frac{A_1}{2} + \frac{5}{4}A_2 + A_3\right)\right]\right\}; \tag{4.13}$$

$$C_{M_0} = \frac{4\pi}{J^2}\left(1 + \sigma + \frac{\gamma^2}{3}\right)K + \mu(C_L\cos\alpha + C_D\sin\alpha) \tag{4.14}$$

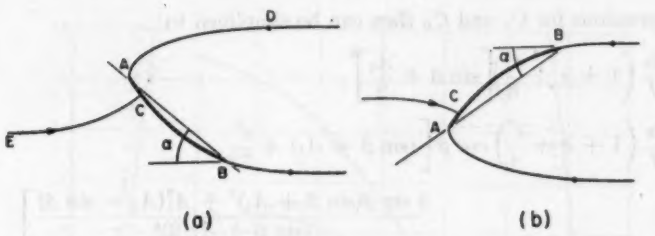


FIG. 4

where β, A_1, A_2, A_3 are given by Eqs. (4.7)–(4.10); J is given by (3.18); K, by (3.19); μ, by (3.20), (3.21) and $\varepsilon = \frac{1}{2}\log(1+\sigma)$.

The results for the flat plate hydrofoil can be deduced from the above expressions by putting $\gamma = 0$. For the circular arc hydrofoil with its convex side toward the approaching stream at positive α (see Fig. 4a), the above formulas still hold if γ assumes a negative value. For flow configurations with positive γ but negative α (see Fig. 4b), the results are that C_D stays the same, but C_L and C_M have oppositive signs as those of flow case (4a).

There are however, several points of complication at which the above results may become invalid, and thus should be applied with great care. First, there in general exists a certain small α, say, α_p, positive or negative, at which the free streamline AD given by Eqs. (3.33), (3.34) would cut into the solid boundary. It could be conjectured that a partial cavitation, that is, with the rear end of the cavity reattached to the boundary, probably is established for α around α_p. The present theory certainly does not cover partially cavitating flows. Second there is another critical value of α, say, $\alpha_c(\gamma, \sigma)$ at which $\beta = 0$, implying that the stagnation point is then at the leading edge. Hence the cavity will shift side on the boundary for $\alpha < \alpha_c$. The value of α_c is in general less than α_p. Third, in flow configurations shown in Fig. 4, the streamline will be in a more critical position at B than at A. In other words, the flow is more likely to separate in front of B for small enough α; and thus a rear part of the boundary will be inside the full cavity. This critical condition will take place when either the slope or the radius of curvature of streamline BD' at B is numerically greater than that of the solid boundary at B. These points will be touched upon in the following explicit calculations, though the clarification of these points requires further study.

a. *Inclined Flat Plate.* For an inclined flat plate, $\gamma = 0$, and hence Eqs. (4.7)–(4.10) reduce to

$$\beta = \alpha + \tfrac{1}{2}[(\alpha^2 + \varepsilon^2)^{\frac{1}{2}} - \alpha] - \tfrac{1}{4}\varepsilon^2/(\pi - \alpha), \tag{4.15}$$

$$A_1 = \tfrac{3}{16}[(\alpha^2 + \varepsilon^2)^{\frac{1}{2}} - \alpha] + \tfrac{3}{32}\varepsilon^2/(\pi - \alpha), \qquad A_2 = 0, \qquad A_3 = -\tfrac{1}{3}A_1 \tag{4.16}$$

and

$$J = 4 + \pi \sin\beta + A_1(\pi + 2.72 \sin\beta). \tag{4.17}$$

The expressions for C_L and C_D then can be simplified to:

$$C_D \cong \frac{2\pi}{J}\left(1+\sigma+\frac{\varepsilon^2}{6}\right)\left[\sin\beta+\frac{A_1}{2}\right]^2, \tag{4.18}$$

$$C_L \cong \frac{2\pi}{J}\left(1+\sigma+\frac{\varepsilon^2}{4}\right)\cos\beta\left[(\sin\beta+A_1)+\frac{\varepsilon^2}{32}\right. \\ \left.\frac{8\sin\beta(\sin\beta+A_1)^2+A_1^2(A_1-\sin\beta)}{(\sin\beta+A_1/2)^4}\right]. \tag{4.19}$$

In the idealized case of $\sigma = 0$ (hence $\varepsilon = 0$), $\beta = \alpha$ and A_1, A_2, A_3 all vanish, the above result then reduces to Rayleigh's theory of the oblique lamina (e.g. ref. 15, p. 102):

$$C_D = \frac{2\pi\sin^2\alpha}{4+\pi\sin\alpha}; \qquad C_L = C_D\cot\alpha. \tag{4.20}$$

When $\alpha = \pi/2$ but $\sigma > 0$, Eqs. (4.17)–(4.19) become

$$C_D \cong \frac{2\pi}{4+\pi}\left(1+\sigma+\frac{1}{6(4+\pi)}\sigma^2\right), \qquad C_L = 0. \tag{4.21}$$

which agree, up to the term in σ, with the known results (see also Eq. (2.2)).

For general values of α and σ, one more physical requirement, however, should be pointed out. For a fully cavitating flow past the oblique flat plate, the local pressure is everywhere normal to the plate and, besides, there is no singular force at leading edge as in the noncavitating case. Consequently C_L and C_D should satisfy the condition

$$C_D/C_L = \tan\alpha \qquad \text{for all } \alpha \text{ and } \sigma. \tag{4.22}$$

In the special cases (i) $\sigma = 0$, (ii) α close to $\pi/2$ and $\sigma > 0$, the above condition is obviously satisfied (see Eqs. (4.20), (4.21)). However, in the general case, it is rather difficult to derive this relation from Eqs. (4.18), (4.19), because of the complicated manner in which the dependence on α and σ appears. Under this circumstance, the condition (4.22) can only be used as a check in numerical computation to assure the correctness of this theory.

Another approximate formula for C_L, when α is small but σ is left arbitrary, has been given by Betz. The main idea is first to linearize the Rayleigh's formula (4.20) to obtain $\pi\alpha/2$ and then by adding to this quantity the pressure coefficient on the upper suction side, namely, σ, to obtain

$$C_L = \tfrac{1}{2}\pi\alpha + \sigma. \tag{4.23}$$

This approximation appears, in general, too rough.

On the other hand, as σ increases the cavity dimension diminishes (see Eqs. (3.37), (3.38)); in some cases it is observed experimentally that the cavity collapses completely for σ approximately greater than 1.5. In the latter flow condition C_L and C_D then resume their noncavitating (aerodynamic) values:

$$C_L \cong 2\pi\sin\alpha, \qquad C_D \cong (2\cos\alpha)C_f \tag{4.24}$$

where C_f is the mean friction coefficient on one side of the plate.

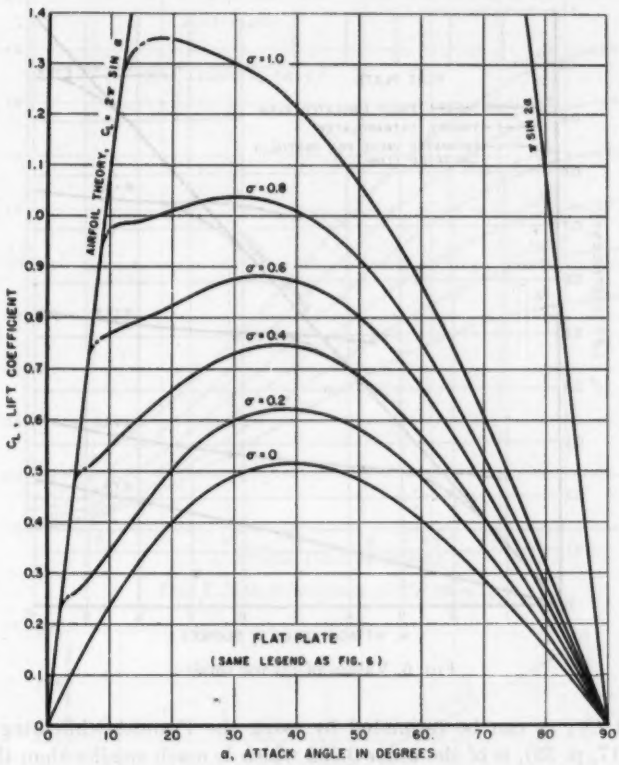


FIG. 5. The dependence of C_L on α

The value of C_L given by (4.19) is plotted against α for different values of σ in Figs. 5 and 6. The aerodynamic value of C_L given by (4.24) is also shown for comparison. For a given value of σ, there is a certain small α, say, α_p at which the cavitating value of C_L becomes equal to the aerodynamic value if the fully cavitating model is assumed still possible (e.g. $\alpha_p = 4.5°$ for $\sigma = 0.4$). A further extrapolation (dotted lines) of Eq. (4.19), without justification, to smaller α would yield an implausible result that the cavitating value of C_L would be greater than its corresponding aerodynamic value. As a physical conjecture, this result is unacceptable. Instead, we expect that near $\alpha = \alpha_p$ partially cavitating flow takes place, a transitional stage between the fully cavitating and fully wetted conditions. This argument is supported by experimental evidence, as shown by the double-dotted lines in Figs. 5 and 6. Thus, the aerodynamic value of C_L for a fully wetted hydrofoil is actually the asymptote to which the cavitating C_L at every σ approaches from below as α decreases from α_p.

The value of C_D given by Eq. (4.18) is similarly plotted in Figs. 7 and 8. In the cavitating range of practical interest, the Reynolds number of the flow is in general very large, say, of the order 5×10^5 or greater. Then the frictional drag

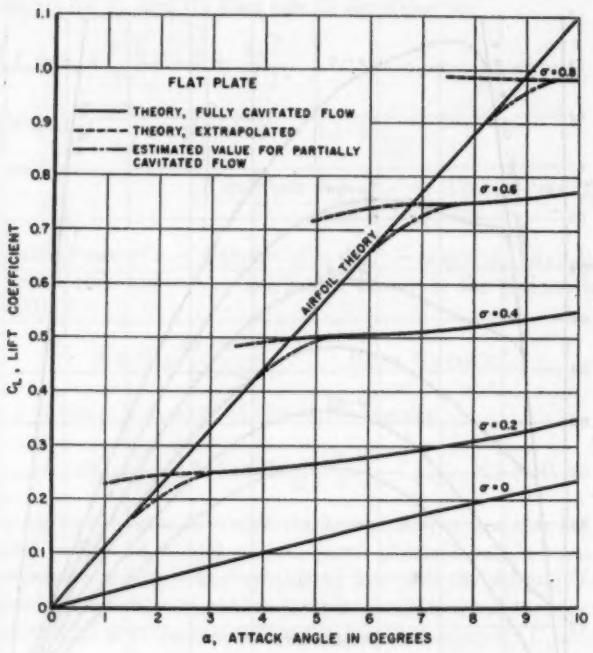


FIG. 6. Values of C_L for small α

coefficient C_f, as can be estimated by using the Prandtl-Schlichting formula (e.g. ref. 17, p. 33), is of the order 0.005 which is much smaller than the cavity C_D for almost all α and σ and can thus be neglected.

A series of experiments (18) was carried out in the Hydrodynamics Laboratory, California Institute of Technology, at about the same time the present theoretical result was obtained. In order to compare the theory with the experiments, C_L and C_D are further cross-plotted against σ in Figs. 9 and 10 in which the experimental data are also shown.* The agreement is very good. As a further check, the value C_L/C_D is plotted against α for several σ and is compared with cot α in Fig. 11. The deviation is less than a few percent, implying the accuracy of the present theory.

Some of the salient points of the previous results may be summarized here.

(i) The results plotted in Figs. 9 and 10 show that for α large, say, greater than 45°, the values of C_L and C_D approach respectively the asymptotes

$$C_L(\sigma, \alpha) = (1 + \sigma)C_L(0, \alpha), \qquad C_D(\sigma, \alpha) = (1 + \sigma)C_D(0, \alpha), \quad (4.25)$$

* In these experimental data, the correction due to tunnel-wall effect was not taken into account. However, the cavitation number σ was computed based on the measured cavity pressure which presumably absorbs a part of the wall-effect correction. For the detailed description, refer to Ref. (18).

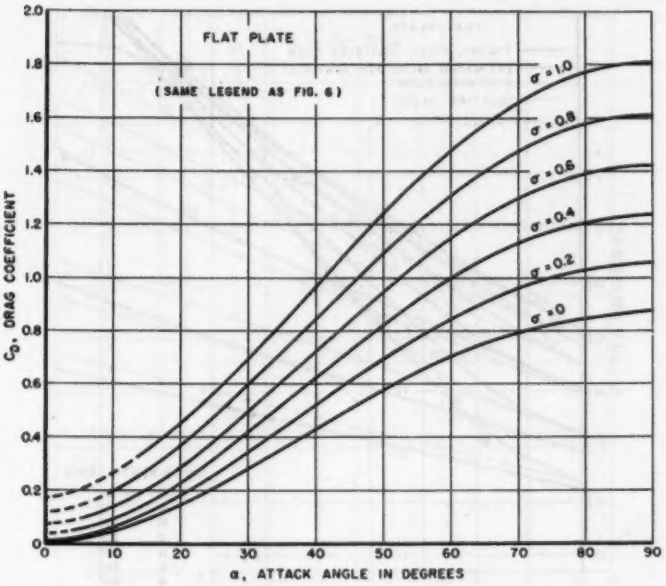


FIG. 7. The dependence of C_D on α

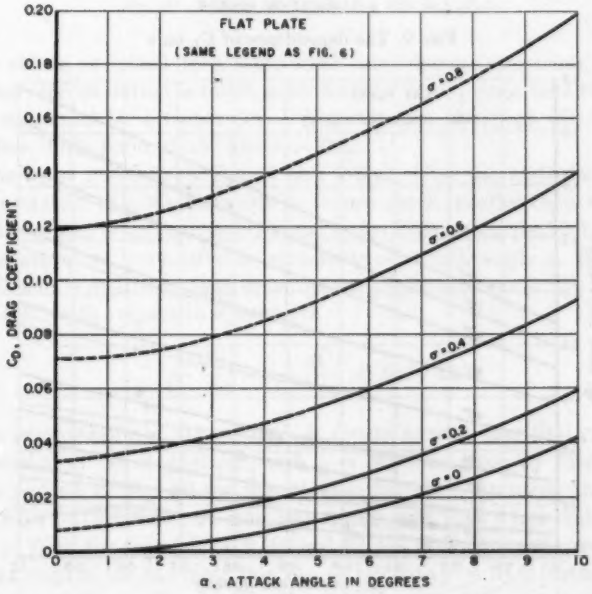


FIG. 8. Values of C_D for small α

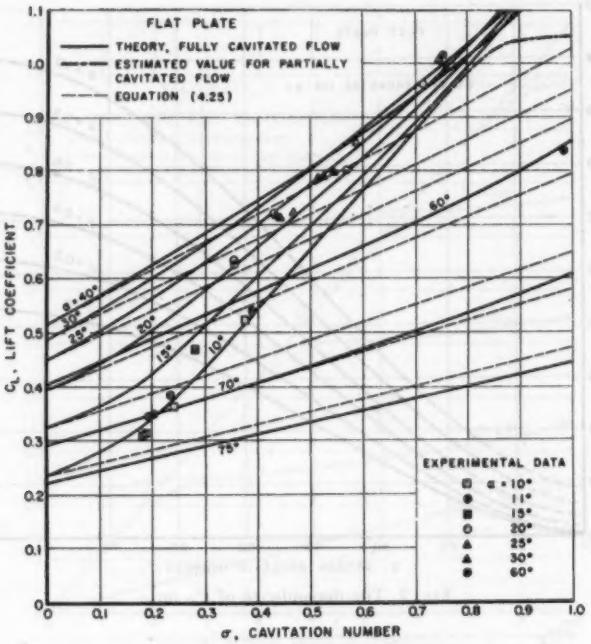


FIG. 9. The dependence of C_L on σ

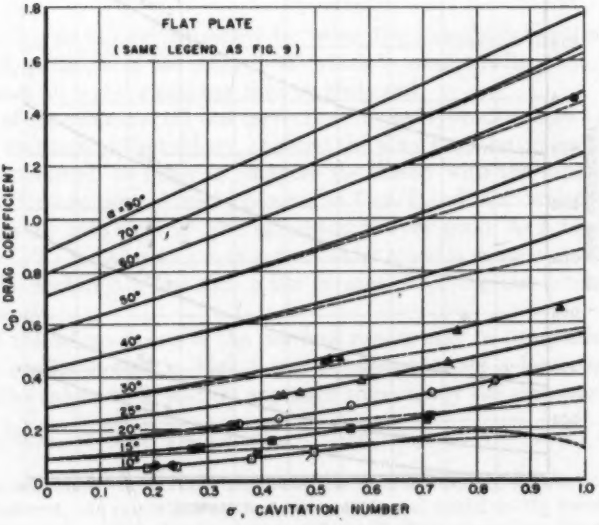


FIG. 10. The dependence of C_D on σ

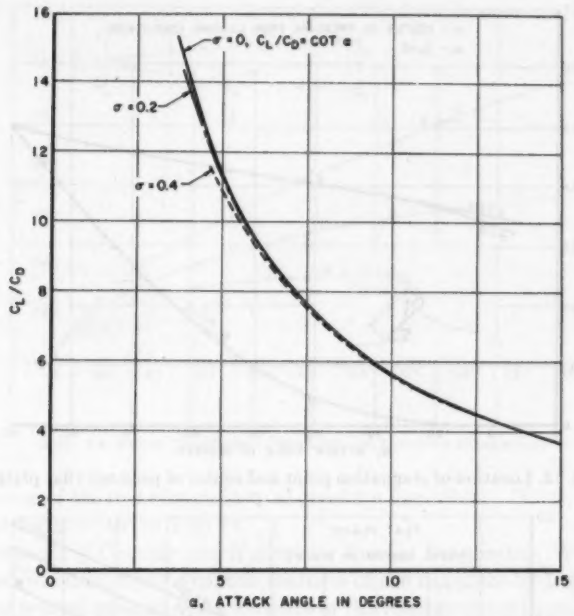


FIG. 11. Values of C_L/C_D for the flat plate

which are shown as dotted lines. For α small, they deviate appreciably from these asymptotes; the deviation is much more marked for C_L since here the deviation of C_D is magnified by a factor cot α. For instance, the slope $dC_L/d\sigma$ becomes greater than unity for $\alpha < 15°$ and $\sigma > 0.2$.

(ii) It is to be noted from Figs. 5 and 6 that, after the hydrofoil at small α is fully cavitated, $dC_L/d\alpha$ decreases to values much smaller than that of fully wetted case, which is approximately equal to 2π. This means that fully cavitating hydrofoils are quite insensitive to variations of attack angle α. However, the drag coefficient is relatively more sensitive due to the relation (4.22). Differentiating Eq. (4.22) with respect to α, we have

$$\frac{1}{C_D}\frac{dC_D}{d\alpha} = \frac{1}{C_L}\frac{dC_L}{d\alpha} + 2 \csc 2\alpha \tag{4.26}$$

Thus, the percentagewise change of C_D is always greater than that of C_L.

The location of the stagnation point, $\mu = S_0/S$, as given by Eqs. (3.18) and (3.20), is plotted in Fig. 12 for this oblique flat plate by using the quantities given in Eqs. (4.15)–(4.17). It is of interest to note that σ has really negligible effect on μ. With the aid of Figs. 9, 10 and 12, the moment coefficient C_{M_0} about the leading edge of the flat plate (see Eq. (4.14) with $\gamma = 0$) is further computed and plotted against σ in Fig. 13. The theoretical value of C_{M_0} is also in fair agree-

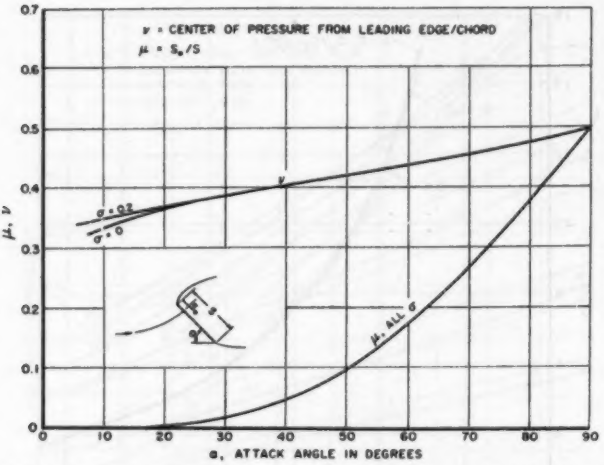


FIG. 12. Location of stagnation point and center of pressure (flat plate)

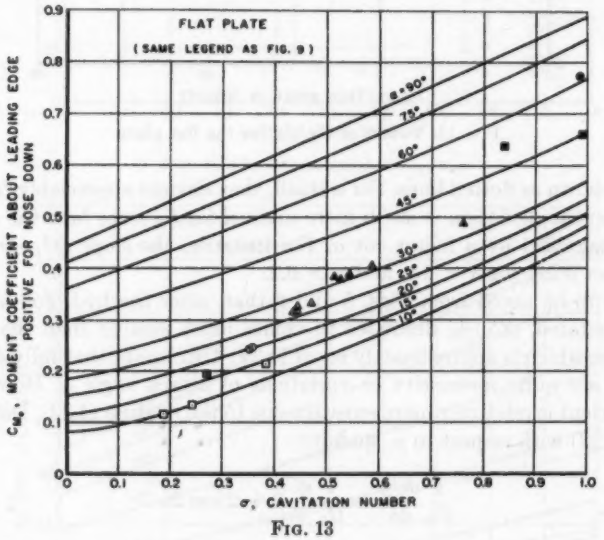


FIG. 13

ment with the experimental data (18). From these results the location of the center of pressure

$$\nu = C_{M_0}/(C_L \cos \alpha + C_D \sin \alpha) \tag{4.27}$$

can be easily deduced. The result shows that ν, like μ, is also independent of σ for all practical range of σ (see Fig. 12). It varies almost linearly from one-third chord at small α to half-chord at $\alpha = \pi/2$.

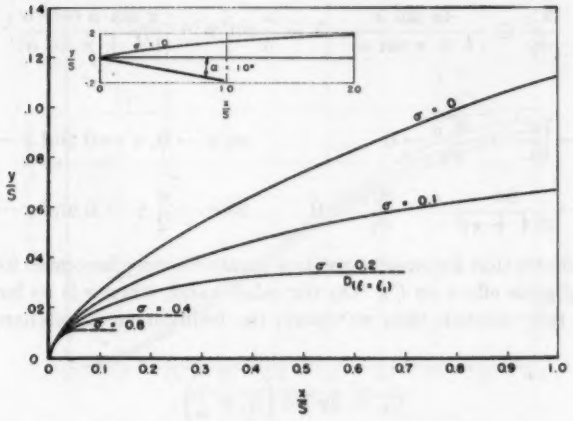


FIG. 14. Some calculated locations of the free streamline

The location of the free streamlines, as given by Eqs. (3.33), (3.34), is computed and plotted for $\alpha = 10°$ in Fig. 14.

b. *An Example of Circular Arc Hydrofoils; Further Discussions.* With a knowledge of some essential hydrodynamic features of the flat plate hydrofoil, we may further note several general characteristics of the camber effect in cavity flows by examining the circular arc profile.

First, when both γ and ε are assumed to be small, then at $\alpha = \pi/2$, we have

$$A_1 \cong \gamma + \frac{9}{8\pi}\varepsilon^2, \qquad A_3 \cong -\frac{\varepsilon^2}{8\pi} - 0.05\gamma, \qquad A_2 \text{ and } (\beta - \alpha) = 0(\gamma\varepsilon^2);$$

and hence

$$C_D \cong \frac{2\pi}{4+\pi}(1+\sigma)\left[1 + \frac{4\gamma}{3(4+\pi)}\right], \quad C_L = 0. \tag{4.28a}$$

Therefore

$$\left(\frac{dC_D}{d\gamma}\right)_{\alpha=\pi/2} \cong \frac{8\pi}{3(4+\pi)^2}(1+\sigma) \cong \frac{1}{6}(1+\sigma). \tag{4.28b}$$

These equations then represent the effect of camber on pure drag problems.

Second, in order to exhibit the effect of camber on C_L and C_D for all α, we consider the limiting case $\varepsilon \to 0$ and $\gamma \to 0$, in which case

$$\frac{dA_1}{d\gamma} \to 1, \qquad \frac{dA_2}{d\gamma} = -\frac{d\beta}{d\gamma} \to \frac{\cos\alpha}{4 + \pi\sin\alpha}.$$

It then follows from Eqs. (4.12), (4.13) that as $\varepsilon \to 0$, $\gamma \to 0$,

$$\frac{dC_L}{d\gamma} \cong \frac{7\pi\cos\alpha}{(4+\pi\sin\alpha)^2}\left\{1 + \frac{16}{21}\sin\alpha + \frac{4}{7}\sin^2\alpha\right\}; \tag{4.29}$$

$$\frac{dC_D}{d\gamma} \cong \frac{4\pi \sin\alpha}{(4+\pi\sin\alpha)^2}\left\{1 - \frac{1}{3}\sin^2\alpha + \frac{\pi \sin\alpha\cos^2\alpha}{4(4+\pi\sin\alpha)}\right\}. \tag{4.30}$$

Therefore,

$$\frac{dC_L}{d\gamma} \to \frac{7\pi}{16}, \quad \frac{dC_D}{d\gamma} \to 0 \qquad \text{as } \alpha \to 0, \gamma \to 0 \text{ and } \varepsilon \to 0, \tag{4.31}$$

$$\frac{dC_D}{d\gamma} \to \frac{8\pi}{3(4+\pi)^2}, \quad \frac{dC_L}{d\gamma} \to 0 \qquad \text{as } \alpha \to \frac{\pi}{2}\ \gamma \to 0 \text{ and } \varepsilon \to 0. \tag{4.32}$$

This result shows that for small α positive camber is very favorable for increasing C_L with negligible effect on C_D. On the other hand, when σ is so large that the hydrofoil is fully wetted, then we obtain the well-known aerodynamic value of C_L:

$$C_L = 2\pi \sin\left(\alpha + \frac{\gamma}{2}\right), \tag{4.33}$$

from which we derive

$$\left(\frac{dC_L}{d\gamma}\right)_{\alpha\to 0,\ \gamma\to 0} = \pi. \tag{4.34}$$

Although the value of $dC_L/d\gamma$ for fully cavitated hydrofoil is less than that in fully wetted flow, a comparison, however, can be made on a different basis. If we compare these values at the same effective angle of atack (aerodynamic)

$$\alpha_e = \alpha + \tfrac{1}{2}\gamma,$$

then the aerodynamic value of C_L is almost the same for same α_e, but the cavitated value of C_L still increases with increase in γ, holding α_e fixed. The rate of increase in this case can be estimated to be $dC_L/d\gamma = \frac{1}{2}$ at equal α_e. The in-

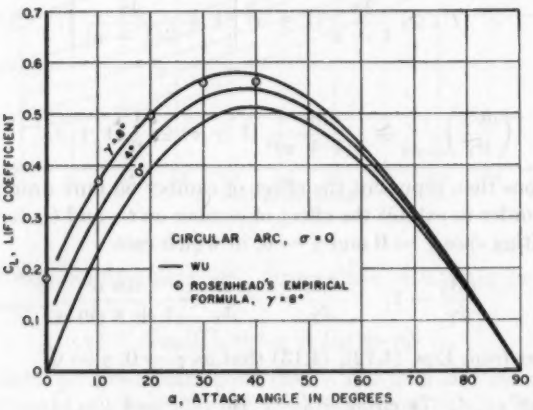


FIG. 15. Effect of camber on C_L at $\sigma = 0$

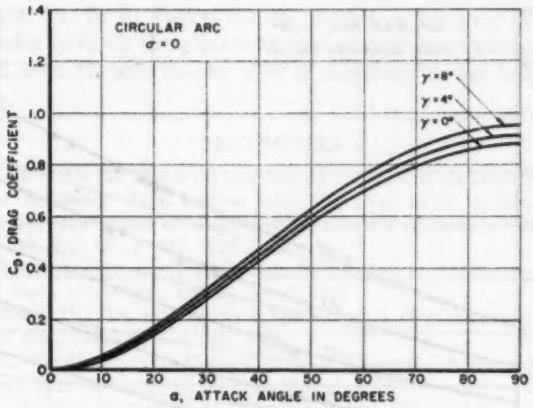


FIG. 16. Effect of camber on C_D at $\sigma = 0$

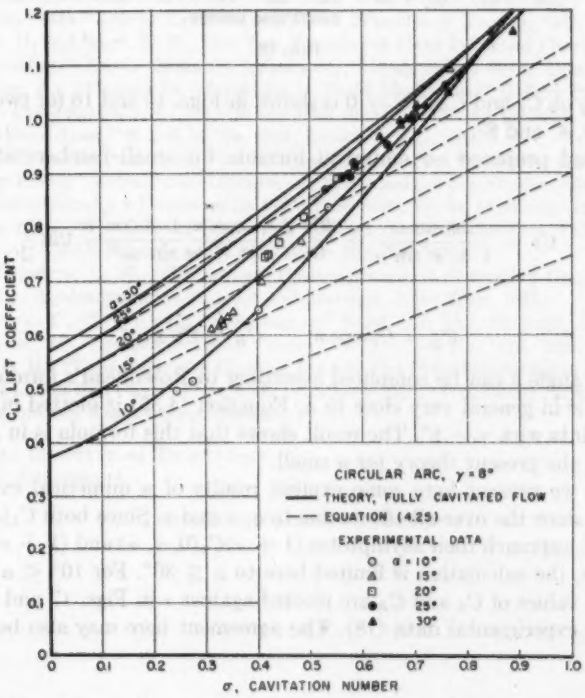


FIG. 17

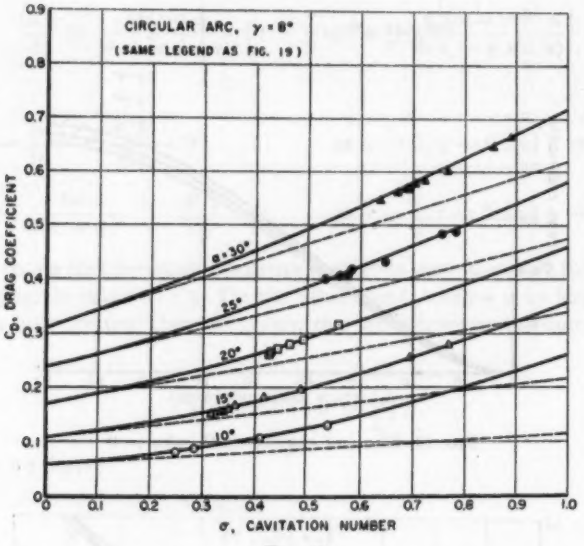


FIG. 18

fluence of γ on C_L and C_D at $\sigma = 0$ is shown in Figs. 15 and 16 for two particular values of γ, 4° and 8°.

Rosenhead proposed an empirical formula for small cambers at $\sigma = 0$ as follows (2):

$$C_P = \frac{2\pi \sin \alpha}{4 + \pi \sin \alpha} + \frac{20\pi}{9} \frac{2 + \cos \alpha + 3 \cos^2 \alpha}{(4 + \pi \sin \alpha)^2} \tan \frac{\gamma}{2} \tag{4.35a}$$

and

$$C_L = C_P \cos \delta, \qquad C_D = C_P \sin \delta \tag{4.35b}$$

where the angle δ can be computed according to Rosenhead's formulation. The value of δ is in general very close to α. Equation (4.35) is plotted in Fig. 15 for several points with $\gamma = 8°$. The result shows that this formula is in good agreement with the present theory for α small.

Finally, we present here some explicit results of a numerical example with $\gamma = 8°$ to show the over-all effects due to σ, α and γ. Since both $C_L(\sigma, \gamma, \alpha)$ and $C_D(\sigma, \gamma, \alpha)$ approach their asymptotes $(1 + \sigma)C_L(0, \gamma, \alpha)$ and $(1 + \sigma)C_D(0, \gamma, \alpha)$ for α large, the calculation is limited here to $\alpha \leqq 30°$. For $10° < \alpha < 30°$, the calculated values of C_L and C_D are plotted against σ in Figs. 17 and 18 together with some experimental data (18). The agreement here may also be considered as good.

Acknowledgments. The author wishes to express his great appreciation for many useful discussions with Professors M. S. Plesset and H. S. Tsien. He also

wishes to thank Dr. B. R. Parkin for his interest in planning an experimental program in order to check the present theory, and for his courtesy in furnishing the data used here. He also thanks Miss Z. Lindberg for her help in numerical computations.

REFERENCES

1. Poole, E. G. C., "On the Discontinuous Motion Produced in an Infinite Stream by Two Plate Obstacles," Proc. London Math. Soc., Vol. 22 (2), p. 425, 1924.
2. Rosenhead, L., "Resistance to a Barrier in the Shape of an Arc of a Circle," Proc. Roy. Soc. (A), Vol. 117, p. 417, 1928.
3. Gilbarg, D. and Serrin, J., "Free Boundaries and Jets in the Theory of Cavitation," J. of Math. and Phys., Vol. 29, No. 1, 1950.
4. Riabouchinsky, D., "On Steady Fluid Motions with Free Surfaces," Proc. London Math. Soc., Vol. 19 (2), p. 206, 1921.
5. Kreisel, G., "Cavitation with Finite Cavitation Numbers," Admiralty Research Laboratory Report, No. R1/H/36, Jan. 1946.
6. Roshko, A., "A New Hodograph for Free-Streamline Theory," NACA TN 3168, July 1954.
7. Eppler, R., "Beiträge zu Theorie und Anwendung der unstetigen Strömungen," Journal of Rational Mechanics and Analysis, Vol. 3, No. 5, p. 591, 1954.
8. Gurevich, M., "Some Remarks on Stationary Schemes for Cavitation Flow about a Flat Plate," David Taylor Model Basin Translation No. 224, 1948.
9. Gilbarg, D. and Rock, H. H., "On Two Theories of Plane Potential Flows with Finite Cavities," Naval Ordnance Laboratory, Memorandum 8718, August 1946.
10. Goldstein, S., "Modern Developments in Fluid Mechanics," Oxford Press, 1950.
11. Fage, A. and Johansen, F. C., "On the Flow of Air Behind an Inclined Flat Plate of Infinite Span," R and M No. 1104, British A.R.C., 1927.
12. Plesset, M. S. and Perry, B., "On the Application of Free Streamline Theory to Cavity Flows," Extrait des Mémoires sur la Mécanique des Fluides, offerts à M. D. Riabouchinsky a l'occasion de son Jubilé Scientifique, Ministère de l'Air, 1954.
13. Reichardt, H., "The Laws of Cavitation Bubbles at Axially Symmetrical Bodies in a Flow," MAP Reports and Translations No. 766, 1946.
14. Milne-Thomson, L. M., "Theoretical Hydrodynamics," Macmillan Co., 1949.
15. Lamb, H., "Hydrodynamics," Dover Publications, New York, 1945.
16. Levi-Civita, T., "Scie e leggi de resistenzia," Rend. Cir. Mat. Palermo, 23, p. 1, 1907.
17. Schlichting, H., "Boundary Layer Theory," NACA TM No. 1218, 1949.
18. Parkin, B. R., "Experiments on Circular Arc and Flat Plate Hydrofoils in Noncavitating and in Full Cavity Flow, California Institute of Technology, Hydrodynamics Laboratory Report (in preparation).

California Institute of Technology
Pasadena, California

(Received September 19, 1955)

THE MAGNETOSTATIC CHARACTERISTICS OF TWO NON-MAGNETIC ELLIPTIC CYLINDERS

By E. E. Jones

1. Introduction. Knowledge of the magnitudes of the mechanical forces between current carrying conductors is desirable from the practical point of view, as has been pointed out by Steinmetz and others, [1]. The application to bus-bar operation has been indicated by Higgins, [2]. The basic type of system is the return circuit of two parallel cylindrical conductors of circular sections, the forces between which are well-known when the materials forming the conductors are non-magnetic in character. It is of interest to determine the deviation from these results when the cylinders have sections which deviate from the circular form: an elliptic section is such a case. It is also possible to deduce an exact expression for the self-inductance of the cylinders, a result which has application in low-frequency transmission line operation. The inductance could presumably be calculated by use of Maxwell's Geometric Mean Distance method, [3], which is merely the vector potential method in another form, but an expression for this inductance has not been located in the literature. The above method involves the determination of the solutions to complicated integrals, and the magnetic properties of the conducting system are obscured. An alternative method is therefore proposed in this paper yielding also the magnetic field at all points inside and outside the conductors, hence the magnetic disturbance due to the presence of such a system of conductors could then be determined.

The technique used follows that originated by Green, [4], in connection with an 'external' type hydrodynamical problem, but the notation has been altered, since the 'internal-external' type magnetostatic problem investigated here takes a simpler form. Although Green's method can be applied directly to the problem of conduction in magnetic cylinders of elliptic section, in view of the fact that the boundary conditions applied at the surfaces of the conductors are complex in form, the theory becomes more involved, c.f. Section 6(3). Fortunately in practice non-ferromagnetic conducting materials, such as copper and aluminium, have magnetic permeabilities which differ but slightly from that of free space. In this case the boundary conditions can be put in a more compact form, and the theory considerably simplified.

It has been decided to employ the rationalized M.K.S.C. or Giorgi system of units, [5], although the theory is not restricted to such a system.

2. The Complex Potential Functions. Two parallel identical cylinders have inductive capacities μ equal to that of free space. The right sections of the cylinders are ellipses with centers at 0 and 0′ respectively, where $00' = 2b$, and their major axes are inclined at an angle 2θ. The regions of space are as marked in Fig. 1, a current I flows in one direction through region (1), and an equal current in the opposite direction through region (3). A right section of the cylinders defines the z-plane, where $z = x + iy$. Rectangular axes, $0x$, $0y$

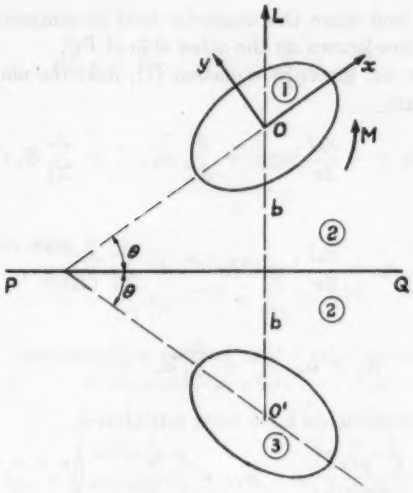


FIG. 1

are taken with origin at 0, in the directions of the semi-major and semi-minor axes respectively of the ellipse defining region (1). The right section intersects the plane bisecting the angle between the inclined major axes of the ellipses in the line PQ, and on this line $z = (x_0 - ib)e^{-i\theta}$, where x_0 is a variable length.

The magnetic field can be expressed in terms of magnetic vector potentials directed parallel to the axes of the cylinders. If $V_s(x, y)$ is the non-zero component of the vector potential in the sth region, where $s = 1, 2, 3$, then complex potential functions $w_s = \phi_s + iV_s$ can be formed, where $\phi_s(x, y)$ is a harmonic function of x and y. With the currents flowing as stated above, the line PQ is a line of force, hence along it V_2 is constant. The required condition to be imposed on w_2 is therefore

$$Im\, w_2 = \text{const.},$$

when $z = (x_0 - ib)e^{-i\theta}$. Another condition exists in region (2), since at large distances from the conducting system, $w_2 \to 0$, the currents having an effect as if they were line currents through 0 and 0′ respectively. These two conditions imposed on w_2 are satisfied if it is assumed that

$$w_2 = \frac{i\mu I}{2\pi} \log \frac{z + 2ibe^{-i\theta}}{z} + \sum_{n=1}^{\infty} \left\{ \frac{A_n}{z^n} + \frac{\bar{A}_n e^{-2in\theta}}{(z + 2ibe^{-i\theta})^n} \right\}, \tag{1}$$

where the A_n are complex constants. In region (2) it is noticed that w_2 is an analytic function of z, hence its imaginary part V_2 is a harmonic function of x and y as required, since in this region no currents flow. The constants A_n are chosen so as to satisfy certain boundary conditions on the ellipse forming the bounding curve to region (1). It is then possible to deduce w_1 and w_2 uniquely

on one side of PQ, and since the magnetic field is symmetrical about PQ the functions w_2 and w_3 are known on the other side of PQ.

The expansion for w_2, given by equation (1), near the ellipse of center 0 can be written in the form

$$w_2 = -\frac{i\mu I}{2\pi}\log z + \sum_{n=0}^{\infty} A_n z^{-n} + \sum_{n=1}^{\infty} B_n z^n, \tag{2}$$

where

$$A_0 = \frac{i\mu I}{2\pi}\log(2ibe^{-i\theta}) + \sum_{n=1}^{\infty}\frac{\bar{A}_n e^{-in\theta}}{(2ib)^n},$$

and

$$B_n = \alpha_{n,0}\, b^{-n} + \sum_{r=1}^{\infty}\alpha_{n,r}\,\bar{A}_r b^{-(n+r)}. \tag{3}$$

Here the following coefficients have been introduced:

$$\alpha_{n,0} = \frac{i^{n-1}\mu I e^{in\theta}}{n\pi 2^{n+1}}, \qquad \alpha_{n,r} = \frac{i^{n-r}e^{i(n-r)\theta}}{2^{n+r}}\binom{n+r-1}{n},$$

where $n \geqq 1$, and

$$\binom{n+r-1}{n} = \frac{(n+r-1)!}{n!(r-1)!}.$$

If the distance between the foci of the ellipse of center 0 is $4c$, then the transformation which maps this ellipse in the z-plane on to the line $\eta = \alpha$, $0 \leqq \xi \leqq 2\pi$, in the ζ-plane, where $\zeta = \xi + i\eta$, is

$$z = c(e^{i\zeta} + e^{-i\zeta}). \tag{4}$$

The semi-major axis of the ellipse is $2c\cosh\alpha$, and the semi-minor axis is $2c\sinh\alpha$. It is assumed that $\eta = +\infty$ corresponds to $z = +\infty$, thus the z-plane outside the line $-2c \leqq x \leqq 2c$ is mapped conformally on to the semi-infinite strip in the ζ-plane given by $0 \leqq \eta < \infty$, $0 \leqq \xi \leqq 2\pi$. Strictly the two-fold Riemann surface in the z-plane with branch points at $z = \pm 2c$ is mapped conformally on to the infinite strip $-\infty < \eta < \infty$, $0 \leqq \xi \leqq 2\pi$, in the ζ-plane, [6], [7].

From equation (4), it can be shown that

$$z^{-n} = \sum_{r=n}^{\infty}\beta_{n,r}e^{ir\zeta},$$

where for $n \geqq 1$,

$$\beta_{n,r}\begin{cases} = 0 & \text{when } r = n+1, n+3, \cdots, \\ = (-1)^{(r-n)/2}\binom{\frac{n+r-2}{2}}{\frac{r-n}{2}}c^{-n}, & \text{when } r = n, n+2, \cdots, \end{cases}$$

and

$$\log z = \log c - i\zeta + \sum_{r=1}^{\infty} \beta_{0,r} e^{ir\zeta},$$

where

$$\beta_{0,r} \begin{cases} = 0, & \text{when } r \text{ is odd,} \\ = \dfrac{2(-1)^{(r/2)+1}}{r}, & \text{when } r \text{ is even.} \end{cases}$$

It can also be shown that

$$z^n = \sum_{r=0}^{\infty} \lambda_{n,r} e^{-ir\zeta} + \sum_{r=1}^{\infty} \gamma_{n,r} e^{ir\zeta},$$

where $\lambda_{n,r}$ and $\gamma_{n,r}$ are zero if n is even and r odd, or vice-versa. Otherwise when $r = n, n-2, \cdots, 0$, (n even), or when $r = n, n-2, \cdots, 1$, (n odd), then

$$\lambda_{n,r} = \binom{n}{\frac{n-r}{2}} c^n, \qquad \gamma_{n,r} = \binom{n}{\frac{n+r}{2}} c^n.$$

Hence substituting for $\log z$, z^n, and z^{-n} in equation (2), and rearranging the order of summation in the double sums, it follows that

$$w_2 = -\frac{\mu I \zeta}{2\pi} + \sum_{n=0}^{\infty} C_n e^{in\zeta} + \sum_{n=1}^{\infty} D_n e^{-in\zeta}, \tag{5}$$

where

$$C_0 = A_0 - \frac{i\mu I}{2\pi} \log c + \sum_{n=1}^{\infty} B_n \gamma_{n,0}, \tag{6}$$

and when $n \geqq 1$, then

$$C_n = -\frac{i\mu I}{2\pi} \beta_{0,n} + \sum_{r=1}^{n} A_r \beta_{r,n} + \sum_{r=n}^{\infty} B_r \gamma_{r,n}, \tag{7}$$

$$D_n = \sum_{r=n}^{\infty} B_r \lambda_{r,n}. \tag{8}$$

In region (1) the non-zero component V_1 of the magnetic vector potential satisfies the Poisson differential equation

$$\nabla^2 V_1 = -\mu J, \tag{9}$$

where J is the uniform current density, such that $I = \pi c^2 J(e^{2\alpha} - e^{-2\alpha})$. If a harmonic function $\psi_1(x, y)$ is introduced defined by

$$V_1 = \psi_1 - \tfrac{1}{4}\mu J z\bar{z}, \tag{10}$$

then equation (9) is satisfied, and if $\Omega_1 = \phi_1 + i\psi_1$, then Ω_1 is an analytic function of z in region (1), and

$$w_1 = \Omega_1 - \tfrac{1}{4} i\mu J z\bar{z}. \tag{11}$$

The magnetic intensity at any point has components H_x, H_y defined by $\mu H_x = \partial V/\partial y$, $\mu H_y = -\partial V/\partial x$. Then in region (1)

$$\mu(H_x - iH_y)_1 = \frac{d\Omega_1}{dz} - \frac{1}{2} i\mu J\bar{z}, \tag{12}$$

and in region (2), where $J = 0$, and $w_2 = \Omega_2$,

$$\mu(H_x - iH_y)_2 = \frac{dw_2}{dz}. \tag{13}$$

In region (1) the magnetic intensity must be finite, single-valued and continuous everywhere. Since $dz/d\zeta = 0$ at the points $z = \pm 2c$ corresponding to $\zeta = 0, \pi$, respectively, which are the foci of the ellipse, then $d\Omega_1/dz$ is finite if if has $dz/d\zeta$ as a factor. A convenient form for the complex potential function in region (1) is therefore

$$\Omega_1 = \sum_{n=1}^{\infty} H_n(e^{in\zeta} + e^{-in\zeta}) + H_0, \tag{14}$$

where H_n, $(n \geqq 0)$, are complex constants. In order for the magnetic intensity to be single-valued and continuous it is necessary that $d\Omega_1/dz$ be continuous across the line of foci of the ellipse. Near points on either side of the line of foci are given by $\zeta = \xi$, and $\zeta = 2\pi - \xi$, hence on looping the points $z = 2c$ or $z = -2c$ once, combinations such as $e^{in\zeta} + e^{-in\zeta}$ return to the same value. It is noticed that such terms occur in the relation for $d\Omega_1/dz$, hence it is single-valued and continuous everywhere in region (1), in particular at all points on the line of foci of the ellipse, which is the line joining the branch points of equation (4).

3. The Boundary Conditions. The boundary conditions to be satisfied on the interface of regions (1) and (2) are

$$V_1 = V_2, \qquad \partial V_1/\partial n = \partial V_2/\partial n. \tag{15}$$

These conditions ensure that the magnetic vector potentials and the tangential components of the magnetic intensities are continuous across the elliptic interface.

From equation (10) the first boundary condition can be written as

$$\psi_1(\xi) - \tfrac{1}{4}\mu J z(\chi)\bar{z}(\bar{\chi}) = V_2(\xi),$$

where $\chi = \xi + i\alpha$, then differentiating with respect to ξ,

$$\frac{\partial \psi_1}{\partial \xi} - \tfrac{1}{4}\mu J z'(\chi)\bar{z}(\bar{\chi}) - \tfrac{1}{4}\mu J z(\chi)\bar{z}'(\bar{\chi}) = \frac{\partial V_2}{\partial \xi}. \tag{16}$$

The second boundary condition can be written as $\partial V_1/\partial \eta = \partial V_2/\partial \eta$, or

$$\frac{\partial \psi_1}{\partial \eta} - \frac{1}{4} i\mu J z'(\chi)\bar{z}(\bar{\chi}) + \frac{1}{4} i\mu J z(\chi)\bar{z}'(\bar{\chi}) = \frac{\partial V_2}{\partial \eta}. \tag{17}$$

But the Cauchy-Riemann conditions for analytic functions give $\partial\psi_1/\partial\eta = \partial\phi_1/\partial\xi$, and $\partial V_2/\partial\eta = \partial\phi_2/\partial\xi$, and these relations in connection with equations (16) and (17) lead to the result

$$\frac{d\Omega_1}{d\zeta} - \frac{1}{2}i\mu J z'(\xi + i\alpha)\bar{z}(\xi - i\alpha) = \frac{dw_2}{d\zeta}, \tag{18}$$

with $\eta = \alpha$. On differentiating the equation

$$\Omega_1 - \tfrac{1}{2}i\mu J \int^{\zeta} z'(\zeta)\bar{z}(\zeta - 2i\alpha)\, d\zeta = w_2 + \text{const.}, \tag{19}$$

with respect to ζ, and substituting $\eta = \alpha$, i.e. $\zeta = \xi + i\alpha$, it is noticed that equation (18) can be reproduced. Thus equation (19) is a relation between the complex potential functions in the adjacent regions (1) and (2), which satisfies identically the boundary conditions on the elliptic profile of the section of the interface. With the transformation of equation (4), it can be shown that equation (19) reduces to

$$w_2 = \Omega_1 - \frac{\mu I}{2\pi} - \frac{1}{4}i\mu J c^2(e^{2i\zeta+2\alpha} + e^{-2i\zeta-2\alpha}) + E + iF, \tag{20}$$

where if $V_1 = V_2$ when $\eta = \alpha$, then

$$F = \frac{\mu\alpha I}{2\pi} - \frac{1}{4}\mu J c^2(e^{2\alpha} + e^{-2\alpha}). \tag{21}$$

The value of E is not recorded here as it is not required in the subsequent analysis.

4. The Determination of the Coefficients. If equations (5) and (14) are substituted into equation (20), then for this relation to be true for all values of ζ,

$$C_0 = H_0 + E + iF, \tag{22}$$

and for $n \geq 1$,

$$C_n = H_n - \tfrac{1}{4}i\mu J c^2 e^{2\alpha}\delta_{n2}, \tag{23}$$

$$D_n = H_n - \tfrac{1}{4}i\mu J c^2 e^{-2\alpha}\delta_{n2}, \tag{24}$$

where

$$\left.\begin{aligned} \delta_{n2} &= 0 \quad \text{if} \quad n \neq 2, \\ &= 1 \quad \text{if} \quad n = 2. \end{aligned}\right\}$$

By eliminating H_n from equations (23) and (24), and expressing the current density J in terms of the total current I, then

$$C_n = D_n - \frac{i\mu I}{4\pi}\delta_{n2}. \tag{25}$$

The coefficient B_n from equation (3) is substituted into equations (7) and (8), and the order of summation is changed, yielding the results

$$C_n = -\frac{i\mu I}{2\pi}\beta_{0,n} + \sum_{r=1}^{n} A_r\beta_{r,n} + \sum_{s=n}^{\infty}\gamma_{s,n}\alpha_{s,0}b^{-s} + \sum_{s=n+1}^{\infty}\sum_{r=n}^{s-1}\gamma_{r,n}\alpha_{r,s-r}\bar{A}_{s-r}b^{-s}, \quad (26)$$

$$D_n = \sum_{s=n}^{\infty}\lambda_{s,n}\alpha_{s,0}b^{-s} + \sum_{s=n+1}^{\infty}\sum_{r=n}^{s-1}\lambda_{r,n}\alpha_{r,s-r}\bar{A}_{s-r}b^{-s}. \quad (27)$$

These expressions are substituted in the conditional relation of equation (25), and on equating coefficients of b^{-s} on each side of the equation, then for all s it follows that

$$\frac{i\mu I}{2\pi}\beta_{0,n} - \sum_{r=1}^{n} A_r\beta_{r,n} = \frac{i\mu I}{4\pi}\delta_{n2}, \quad (n \geqq 1), \quad (28)$$

$$\left.\begin{aligned}\gamma_{s,n}\alpha_{s,0} + \sum_{r=n}^{s-1}\gamma_{r,n}\alpha_{r,s-r}\bar{A}_{s-r} &= \lambda_{s,n}\alpha_{s,0} + \sum_{r=n}^{s-1}\lambda_{r,n}\alpha_{r,s-r}\bar{A}_{s-r}, \quad (s > n \geqq 2),\\ &= \lambda_{n,n}\alpha_{n,0}, \quad (n = s \geqq 2),\\ &= 0, \quad (n > s \geqq 2).\end{aligned}\right\} \quad (29)$$

Using the definitions of the coefficients $\beta_{r,n}$, $\alpha_{r,n}$, $\lambda_{r,n}$, and $\gamma_{r,s}$, as defined in Section 2, it can be readily shown that only equation (28) gives values for A_n, the equation (29) consisting of identities. It can be proved from equation (28) that the coefficients A_n are all purely imaginary, and are zero when n is odd. When n is even, the coefficients are in general non-zero, the first few coefficients being

$$A_2 = i\mu Ic^2/4\pi, \qquad A_4 = i\mu Ic^4/4\pi,$$

$$A_6 = 5i\mu Ic^6/12\pi, \qquad A_8 = 7i\mu Ic^8/8\pi.$$

Having determined the coefficients A_n it is now possible to deduce the magnetic field intensities both inside and outside the conducting cylinders.

5. The Forces acting on the Cylinders. The force **F** per unit length of the cylinder is given by

$$\mathbf{F} = \mu\int_c \{\mathbf{H}(\mathbf{H}\cdot\mathbf{n}) - \tfrac{1}{2}H^2\mathbf{n}\}\,ds,$$

where **H** is the magnetic intensity external to the curve c, which is drawn in the right section of the cylinder, but external to the cylinder, in such a way that it does not displace any singularities of the external magnetic field, [5], [8]. The unit outward drawn normal to the curve c in its plane is **n**. If **H** has axial components H_x, H_y, and **n** has axial components $n_x = dy/ds$, $n_y = -dx/ds$, then the axial components X, Y of the force F can be deduced from

$$Y + iX = \frac{1}{2\mu}\int_c \left(\frac{dw_2}{dz}\right)^2 dz. \quad (30)$$

This integral gives the force components acting on the cylinder enclosing region

(1) if w_2 is interpreted as defined by equation (2). On substituting from equation (2), the residues of the integral are easily evaluated to give

$$Y + iX = IB_1 - \frac{2i\pi}{\mu} \sum_{n=1}^{\infty} n(n+1)A_n B_{n+1}. \tag{31}$$

If L is the force component in the direction 0'0 of Fig. 1, and D is the component along PQ, then $L - iD = (Y + iX)e^{-i\theta}$, and it can readily be shown that $D = 0$, and

$$L = \frac{\mu I^2}{4\pi b} + \sum_{n=1}^{\infty} \frac{nIA_{2n}\cos 2n\theta}{2^{2n-1}(ib)^{2n+1}} - \sum_{n=1}^{\infty} \frac{\pi(4n)!A_{2n}^2}{\mu\{(2n-1)!\}^2 2^{4n} b^{4n+1}} + \sum_{n=1}^{\infty}\sum_{r=n+1}^{\infty} \frac{\pi(2n+2r)!A_{2r}A_{2n}(-1)^{n-r+1}\cos 2(n-r)\theta}{\mu(2n-1)!(2r-1)!\,2^{2n+2r-1}b^{2n+2r+1}}. \tag{32}$$

To the order of approximation which includes terms involving b^{-9}, the resultant force is given by the expression

$$\frac{\pi L}{\mu I^2} = \frac{1}{4b} - \frac{c^2}{8b^3}\cos 2\theta + \frac{c^4}{32b^5}(3 + 2\cos 4\theta) - \frac{5c^6}{128b^7}(6\cos 2\theta + \cos 6\theta) + \frac{7c^8}{256b^9}(10 + 10\cos 4\theta + \cos 8\theta) + \cdots. \tag{33}$$

The vector movement $\mathbf{M}$ of the couple acting in an anticlockwise sense on a cylinder is given by

$$\mathbf{M} = \mu \int_c \{(\mathbf{H}\cdot\mathbf{n})(\mathbf{r}\times\mathbf{H}) - \tfrac{1}{2}H^2(\mathbf{r}\times\mathbf{n})\}\, ds,$$

where $\mathbf{r}$ is the radius vector drawn from the axis to a point on the curve c which encloses the cylinder. If $\mathbf{r}$ has components x, y, then as in the case of the determination of the force the moment of the couple is given by

$$M = \mathrm{Re}\,\frac{1}{2\mu}\int_c \left(\frac{dw_2}{dz}\right)^2 z\, dz. \tag{34}$$

If the cylinder is that enclosing region (1), then

$$M = \mathrm{Re}\,-\frac{2\pi i}{\mu}\sum_{n=1}^{\infty} n^2 A_n B_n, \tag{35}$$

and on substituting from equation (3), then

$$M = \sum_{n=1}^{\infty} \frac{nIi^{2n+3}A_{2n}\sin 2n\theta}{2^{2n-1}b^{2n}} + \sum_{n=1}^{\infty}\sum_{r=n+1}^{\infty} \frac{\pi(n-r)(2n+2r-1)!A_{2n}A_{2r}(-1)^{n-r+1}\sin 2(n-r)\theta}{\mu(2n-1)!(2r-1)!\,2^{2n+2r-2}b^{2n+2r}}. \tag{36}$$

To the order of approximation which includes a term involving b^{-8}, the moment of the couple is given by the expression

$$\frac{\pi M}{\mu I^2} = -\frac{c^2}{8b^2}\sin 2\theta + \frac{c^4}{16b^4}\sin 4\theta - \frac{5c^6}{128b^6}(2\sin 2\theta + \sin 6\theta) + \frac{7c^8}{256b^8}(5\sin 4\theta + \sin 8\theta) + \cdots . \tag{37}$$

Since there is symmetry about the line PQ of Fig. 1, the force and couple acting on the cylinder with axis through $0'$ are equal in magnitude but opposite in direction to the force and couple respectively acting on the cylinder with axis through 0.

It is noticed that the magnitudes of L and M are dependent only on the length c when the cylinders are inclined at a fixed angle 2θ at a fixed distance $2b$ apart. The force and moment of the couple are thus the same for all cylinders of the same confocal elliptic system under the above conditions. This system of conductors includes the case of two inclined strip conductors of length $4c$. Table I illustrates the variation in the function $F = \pi cL/\mu I^2$ as θ and the ratio $d = b/2c$ varies.

TABLE I

The Variation of F

θ	$d = 1$	1.5	2	2.5	3	4	5
0°	0.113	0.079	0.061	0.049	0.041	0.031	0.025
45°	0.126	0.083	0.063	0.050	0.042	0.031	0.025
90°	0.149	0.089	0.065	0.051	0.042	0.031	0.025

The force L between the cylinders is a repulsive force, decreasing in magnitude as the separation of the cylinders increases for any given value of θ, and as the angle θ decreases in the range $0 \leqq \theta \leqq 90°$ for any given value of d. The decrease is greater at smaller values of d for constant θ, and at larger values of θ for constant d. The effect of a change in the angle of inclination of the cylinders is negligible when $b > 8c$.

When the cylinders are of circular sections the force between the cylinders is $L_0 = \mu I^2/4\pi b$, and the deviation from the force on the elliptic cylinders can be measured by L/L_0. The variation of L/L_0 as the ratio $2c/b$ varies, with values of θ given by 0°, 45°, 90°, respectively, is illustrated in Fig. 2. For a given cylinder, or all cylinders in the same confocal system, $L > L_0$ for $\theta = 90°$, and $L < L_0$ for $\theta = 0°$, the variation of L/L_0 being small for $\theta = 45°$. For a given value of θ the force $L \to L_0$ as the cylinders separate, the difference $|L - L_0|$ being greater at values of θ near 90° than at θ near zero.

The moment of the couple about the center of each cylinder is zero when $\theta = 0°$ and when $\theta = 90°$, and when $0° < \theta < 90°$ the couple tends to orientate the cylinders so that the major axes of their sections become parallel. When $\theta = 45°$ the values of $G = -\pi M/\mu I^2$ as $d = b/2c$ varies are given by Table II.

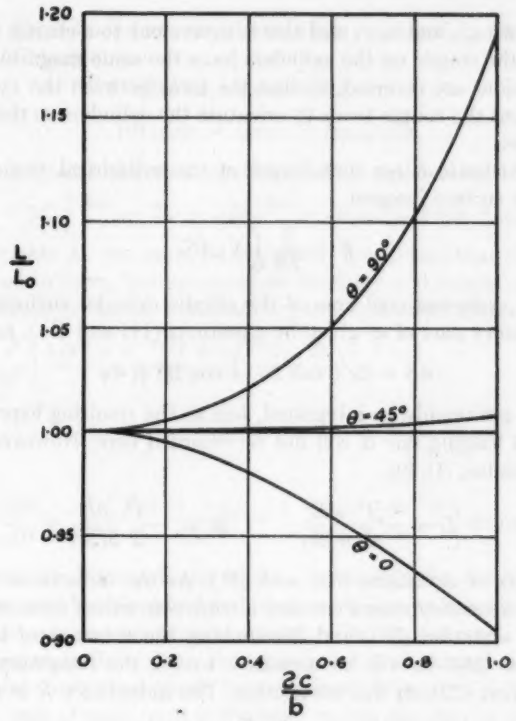


FIG. 2

TABLE II

The variation of G

d =	1	1.5	2	2.5	3	4	5
G =	0.0319	0.0139	0.0078	0.0050	0.0035	0.0019	0.0013

The moment of the couple decreases as the cylinders separate, the rate of decrease being greater at smaller values of d.

6. Further Considerations. (i). If the currents in the cylinders flow in the same direction, the line PQ of Fig. 1 intersects the lines of magnetic force orthogonally. Thus in region (2) it is necessary that Re w_2 = const., when $z = (x_0 - ib)e^{-i\theta}$, and at large distances from the cylinders $w_2 \to -(i\mu I/\pi) \log z$. Such a complex potential function is

$$w_2 = -\frac{i\mu I}{2\pi} \log z(z + 2ibe^{-i\theta}) + \sum_{n=1}^{\infty} \left\{ \frac{A_n}{z^n} - \frac{\bar{A}_n e^{-2in\theta}}{(z + 2ibe^{-i\theta})^n} \right\}.$$

The theory is developed as before, the only difference lies in a change in sign

in the coefficients $\alpha_{n,0}$ and $\alpha_{n,r}$, and this is equivalent to a change in sign of B_n. The force and the couple on the cylinders have the same magnitudes as before, but their directions are reversed, so that the force between the cylinders is one of attraction, and the couple tends to orientate the cylinders so that their minor axes are parallel.

(*ii*). The inductance R per unit length of the cylindrical section, [3], is deduced from the surface integral

$$R = \frac{2}{IS}\int_S V_1\, dS, \tag{38}$$

where S is the cross-sectional area of the elliptic cylinder enclosing region (1), V_1 is the imaginary part of w_1 given by equations (11) and (14), and

$$dS = 2c^2(\cosh 2\eta - \cos 2\xi)\, d\xi\, d\eta.$$

Equation (38) can readily be integrated, but as the resulting expression for the inductance is a lengthy one it will not be recorded here. However the application of the formulae, [1], [9],

$$L = -\frac{I^2}{2}\frac{\partial R}{\partial(2b)}, \qquad M = -\frac{I^2}{2}\frac{\partial R}{\partial(2\theta)} \tag{39}$$

yield the results of equations (33) and (37). As the inductance formula is of interest in low frequency return circuits, a more convenient form can be deduced by integrating equation (39), and determining the constant of integration by use of equation (38). It will be necessary to use the imaginary part of H_0, given by equation (22), in this connection. The inductance R is then given by the relation

$$\frac{\pi R}{\mu} = \alpha - \frac{1}{4} - \log\frac{2b}{c} + \sum_{n=1}^{\infty}\frac{\pi A_{2n}\cos 2n\theta}{\mu I i^{2n+1}2^{2n-2}b^{2n}} + \sum_{n=1}^{\infty}\frac{\pi^2(4n-1)\,!A_{2n}^2}{\mu^2 I^2\{(2n-1)\,!\}^2 2^{4n-2}b^{4n}}$$
$$+ \sum_{n=1}^{\infty}\sum_{r=n+1}^{\infty}\frac{\pi^2(2n+2r-1)\,!A_{2r}A_{2n}(-1)^{n-r+1}\cos 2(n-r)\theta}{\mu^2 I^2(2n-1)\,!(2r-1)\,!2^{2n+2r-2}b^{2n+2r}},$$

and to the order of approximation which includes terms containing b^{-8}, it follows that

$$\frac{\pi R}{\mu} = \alpha - \frac{1}{4} - \log\frac{2b}{c} - \frac{c^2}{4b^2}\cos 2\theta + \frac{c^4}{32b^4}(3 + 2\cos 4\theta)$$
$$- \frac{5c^6}{192b^6}(6\cos 2\theta + \cos 6\theta) + \frac{7c^8}{512b^8}(10 + 10\cos 4\theta + \cos 8\theta) + \cdots.$$

It is noticed that the inductance depends on the lengths of the semi-axes of the ellipse, since it is a function of α.

(*iii*). For a detailed investigation of the problem considered in this paper it is necessary to assume that the permeabilities μ_1/μ_2, and μ_3/μ_2 of the two cylinders are equal to each other, but different from unity. The theory originated by Green, [4], can then be applied directly to the complex potential functions of equations (5) and (14), with the boundary conditions $V_1 = V_2$, $\mu_1\,\partial V_2/\partial n = \mu_2\,\partial V_1/\partial n$,

when $\eta = \alpha$. These boundary conditions when expressed in terms of the complex potentials become

$$\mathrm{Im}\,(w_1 - w_2) = 0$$

$$\mathrm{Re}\,(\mu_2 w_1 - \mu_1 w_2) = I\mu_1\mu_2\zeta/2\pi.$$

If $\lambda = (\mu_1 - \mu_2)/(\mu_1 + \mu_2)$, then equation (25) is replaced by

$$C_n - \lambda C_n e^{-2n\alpha} = D_n - \lambda \bar{D}_n e^{2n\alpha} - i\mu_1\mu_2 I\delta_{n2}/2\pi(\mu_1 + \mu_2),$$

and the coefficients A_n can again be evaluated. It is found that those coefficients are complicated in form, but nevertheless the force and couple on the cylinders can be determined as before. For example, to a first order approximation

$$\frac{\pi L}{\mu_2 I^2} = \frac{1}{4b} - \frac{c^2}{8b^3}\left[\frac{\lambda(e^{2\alpha} - e^{-2\alpha})(1 + \lambda e^{-2\alpha}\cos 2\theta)}{1 - \lambda^2 e^{-4\alpha}} + \frac{\{1 - \lambda(1 - 2e^{-4\alpha})\}\cos 2\theta}{1 + \lambda e^{-4\alpha}}\right], \tag{40}$$

and

$$\frac{\pi M}{\mu_2 I^2} = \frac{c^2}{8b^3}\left[\frac{\lambda^2(e^{-4\alpha} - 1)}{1 - \lambda^2 e^{-4\alpha}} - \frac{1 - \lambda(1 - 2e^{-4\alpha})}{1 + \lambda e^{-4\alpha}}\right]\sin 2\theta, \tag{41}$$

are respectively the force and the couple on the cylinder enclosing region (1). The inductance can be calculated as before, with new values for the coefficients A_n, but in this case the relations of equation (39) do not hold.

When the currents flow along the cylinders in the same direction the force and couple differ from those given by equations (40) and (41). In fact in these equations the sign of each term is changed, except the sign of the first term in the square brackets.

REFERENCES

1. C. P. Steinmetz. Mechanical Forces in Magnetic Fields. A.I.E.E. Trans. 30, 367–85, (1911). Discussion, p. 392–3.
2. T. J. Higgins. The Design of Bus-Bar Industrial Distribution Systems: an Epitomization of Available Data. A.I.E.E. Trans. 64, 385–400, (1945).
3. C. Snow. Formulas for Computing Capacitance and Inductance. Nat. Bur. Stands. Circular 544, p. 64, (1954).
4. A. E. Green. The Two-Dimensional Aerofoil in a Bounded Stream. Quart. J. Math. (Oxf. Series), 18, 167–177, (1947).
5. J. A. Stratton. Electromagnetic Theory. (McGraw Hill Book Co. Inc., New York, 1941), pp. 16, 155.
6. R. M. Morris. The Internal Problems of Two-Dimensional Potential Theory. Math. Ann. 116, 374–400, (1939).
7. ——— The Boundary-value Problems of Plane Stress. Quart. J. Mech. and Appl. Math. 4, 248–256, (1951).
8. F. B. Brown, Jr. Electric and Magnetic Forces: A Direct Calculation, I. Amer. J. Phys. 19, 290–304, (1951).
9. B. Hague. Electromagnetic Problems in Electrical Engineering. (Oxf. Univ. Press, 1929), p. 139.

University of Nottingham, England

(Received September 23, 1955)

ANALYSIS OF A GENERAL SYSTEM FOR THE DETECTION OF AMPLITUDE-MODULATED NOISE*

By Emanuel Parzen and Norman Shiren

1. Introduction. The problem of the detection of signals having the character of amplitude-modulated random noise is of recent interest in communications theory. For example, a paper by Deutsch [1] treats the effect of a linear detector on such signals when the modulation is periodic.

In this paper, we treat a general detection system involving square-law detectors. We shall compute various statistics of the output, with and without modulation, and, from these, compute a detection criterion which we define as a measure of the detectability of the modulated noise.

The system is diagrammed in Fig. 1. The input filters H_1 and H_2 include the characteristics of any receiving equipment which might ordinarily precede them. The only assumptions made on the filters K_1 and K_2 are that they do not pass *dc*. The output integrator is perfect with integration time T.

We take the input $u(t)$ to be the sum of a stationary Gaussian noise $y(t)$ modulated with index m by a modulating function $g(t)$, which may be random or not, and a stationary Gaussian noise $z(t)$ which represents the background noise. Thus

$$u(t) = y(t)[1 + mg(t)] + z(t) \tag{1.1}$$

We recall a few definitions of mathematical noise theory. A time function $y(t)$ is said to be a noise, or a random time function, if $y(t)$ contains, in addition to t, random parameters, so that at each time t, $y(t)$ does not have a definite value but only an ensemble of possible values which it assumes in accordance with some given probability distribution. The ensemble average, denoted by $\langle \cdots \rangle$, is an average taken with respect to these random parameters. The time-ensemble average, denoted by $\langle \cdots \rangle_{TAv}$, is an average over time of an ensemble average; thus

$$\langle y(t) \rangle_{TAv} = \lim_{T \to \infty} \frac{1}{T} \int_0^T \langle y(t) \rangle \, dt \tag{1.2}$$

The auto-correlation function $R_y(\tau)$ of noise is, in our opinion, best defined by the time-ensemble average

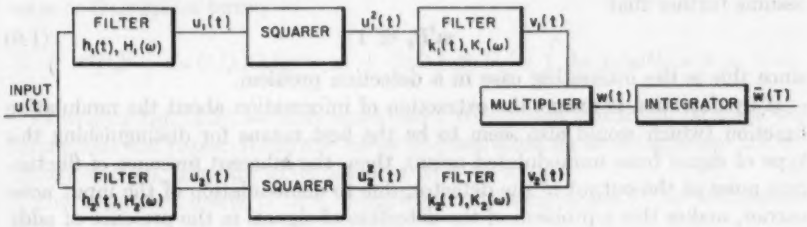
$$R_y(\tau) = \langle y(t)y(t + \tau) \rangle_{TAv} = \lim_{T \to \infty} \frac{1}{T} \int_0^T \langle y(t)y(t + \tau) \rangle \, dt \tag{1.3}$$

The power spectrum $G_y(w)$, which is a measure of the frequency distribution of the average total power

$$P_y = \langle y^2(t) \rangle_{TAv} \tag{1.4}$$

* This work was supported by the Office of Naval Research under contract N6-ONR-27135. Reproduction in whole or in part is permitted for any purpose of the United States Government.

A GENERAL DETECTION SYSTEM FOR AMPLITUDE MODULATED NOISE



$$u_j(t) = \int h_j(\alpha)\ u(t-\alpha)\, d\alpha \qquad j = 1,2$$

$$v_j(t) = \int k_j(\xi)\ u_j^2(t-\xi)\, d\xi \qquad j = 1,2$$

$$w(t) = v_1(t)\ v_2(t)$$

$$\bar{w}(T) = \int_0^T w(t)\, dt$$

FIRST FILTERS HAVE IMPULSE TRANSFER FUNCTIONS $h_j(t)$,
FREQUENCY RESPONSE FUNCTIONS $H_j(\omega) = \int e^{it\omega} h_j(t)\, dt$
SECOND FILTERS HAVE IMPULSE TRANSFER FUNCTIONS $k_j(t)$,
FREQUENCY RESPONSE FUNCTIONS $K_j(\omega) = \int e^{it\omega} k_j(t)\, dt$
AND DO NOT PASS D.C., $K_j(0) = 0$

FIG. 1

is related to $R_y(\tau)$ by the Fourier transformation (Wiener-Khintchine theorem)

$$G_y(\omega) = \frac{2}{\pi} \int_0^\infty \cos \tau\omega R_y(\tau)\, d\tau \tag{1.5}$$

$$R_y(\tau) = \int_0^\infty \cos \tau\omega G_y(\omega)\, d\omega \tag{1.6}$$

A noise $y(t)$ is said to be *stationary* if the ensemble average $\langle y(t)y(t+\tau)\rangle$ depends only on the time difference τ, and not on the initial time t. For stationary noises, then, the distinction between ensemble averages and time-ensemble averages disappears.

We make the following assumptions.

Both noise functions $y(t)$ and $z(t)$ and the modulating function $g(t)$ are taken to have zero mean level, i.e.

$$\langle y(t)\rangle = \langle z(t)\rangle = \langle g(t)\rangle_{TAv} = 0 \tag{1.7}$$

It then follows that all higher odd moments of $y(t)$ and $z(t)$ also vanish.

We assume also that third moments (in the sense of time-ensemble averages) of $g(t)$ vanish. This assumption is justified in many communication applications in which the modulating functions are periodic in time and possess zero time averages. It is also justified for random modulating functions with symmetric statistics.

In addition, the functions $y(t)$, $z(t)$ and $g(t)$ are taken to be statistically independent of each other. Then the total input power P_u is given by,

$$P_u = \langle u^2(t)\rangle_{TAv} = P_y(1 + m^2P_g) + P_z \tag{1.8}$$

Thus m^2P_g is the ratio of the sideband power to the carrier power. We shall assume further that

$$m^2P_g \ll 1 \tag{1.9}$$

since this is the interesting case in a detection problem.

If by detection is meant the extraction of information about the modulating function (which would also seem to be the best means for distinguishing this type of signal from unmodulated noise), then, the inherent presence of fluctuation noise at the output of the detector, due to demodulation of the input noise carrier, makes this a problem of the detection of signals in the presence of additive noise.

In order to see this more clearly, and to better understand the action of the general system of Fig. 1, we shall compute the output power spectrum of a square-law detector when the input is modulated noise. This calculation also gives a simple illustration of the methods used in treating the general system.

2. Power Spectrum after a Square-Law Detector. Let the input of a square-law detector be given by (1.6), with $z(t) = 0$, so that for the present we assume that there is no background noise. Then, the output $v(t)$ of the squarer is given by

$$v(t) = u^2(t) = y^2(t)[1 + 2\,mg(t) + m^2g^2(t)]. \tag{2.1}$$

In order to compute the power spectrum $G_v(\omega)$ of $v(t)$, we first compute the autocorrelation $R_v(\tau)$, and then obtain $G_v(\omega)$ by means of the Wiener-Khintchine theorem.

$$\begin{aligned} R_v(\tau) &= \langle v(t)v(t+\tau)\rangle_{TAv} = \langle y^2(t)y^2(t+\tau)\rangle\{1 + m^2[\langle g^2(t)\rangle_{TAv} \\ &\quad + \langle g^2(t+\tau)\rangle_{TAv} + 4\langle g(t)g(t+\tau)\rangle_{TAv}] + m^4\langle g^2(t)g^2(t+\tau)\rangle_{TAv}\} \end{aligned} \tag{2.2}$$

The terms in (2.2) which would involve m and m^3 vanish, since we have assumed that the first- and third-order time-ensemble averages of $g(t)$ vanish.

To compute the foregoing fourth order ensemble average of $y(t)$, we use the formula given in Appendix A, which expresses the higher order moments of a Gaussian random time function in terms of its autocorrelation function $R_y(\tau)$. We have

$$\langle y^2(t)y^2(t+\tau)\rangle = R_y^2(0) + 2R_y^2(\tau). \tag{2.3}$$

As regards the fourth order time-ensemble average of $g(t)$, we make the assumption (which is fulfilled by a large number of random time functions) that

$$m^4\langle g^2(t)g^2(t+\tau)\rangle_{TAv} \ll m^2R_g(0) \tag{2.4}$$

We may henceforth ignore the term in (2.2) involving m^4.

The power spectrum $G_v(\omega)$ is now easily obtained by taking the Fourier integral of (2.2). The resulting terms group themselves naturally into three groups, where we write P_y and P_g for $R_y(0)$ and $R_g(0)$ respectively.

The power spectrum $G_v(\omega)$ is the sum of:

dc terms, $$2\delta(\omega)P_y^2[1 + m^2P_g]^2 \tag{2.5}$$

steady state or signal terms, $4m^2P_y^2G_g(\omega)$,
noise or fluctuation terms,

$$[1 + m^2P_g]^2 \int d\mu\, G_y(\mu)G_y(\omega - \mu) + 2m^2 \int d\nu\, G_g(\nu) \int d\mu\, G_y(\mu)G_y(\mu + \omega - \nu).$$

In writing Eq. (2.5), we have employed a convention which will be employed throughout the paper. When the limits of integration on an integral are not explicitly written, they are to be taken to be $-\infty$ to ∞.

The integrated power is then, dropping terms involving m^2P_g by assumption (1.9),

$$\begin{array}{ll} dc & P_y^2 \\ \text{signal} & 4m^2P_y^2P_g \\ \text{noise} & 2P_y^2. \end{array} \tag{2.6}$$

Thus, as stated at the end of Sec. 1, at the output of the square-law detector the problem reduces to that of detecting a signal of total power $4m^2P_y^2P_g$ in the presence of an additive noise of total power $2P_y^2$.

This problem is usually attacked by analyzing the detector output with a band-pass or low-pass filter followed by squaring and averaging to determine the mean square signal and noise.

It is easily seen that this detection system, consisting of receiver, square-law detector, filter, squarer, and finite time integrator is a special case of the system of Fig. 1, obtained by setting $H = H_1 = H_2$ and $K = K_1 = K_2$. This case is treated in detail in Sec. 4.

3. The Detection Criterion. Suppose that it is possible to measure the output $w(T)$ of the system of Fig. 1 when the input is unmodulated noise. A large number of such measurements will have an average and a mean square fluctuation given by the ensemble average $\langle w(T)\rangle_n$ and the variance

$$\sigma_n^2 = \langle w^2(T)\rangle_n - \{\langle w(T)\rangle_n\}^2 \tag{3.1}$$

Similarly, a large number of measurements of the output when the input is amplitude-modulated noise will have mean value $\langle w(T)\rangle_{mn}$ and variance

$$\sigma_{mn}^2 = \langle w^2(T)\rangle_{mn} - \{\langle w(T)\rangle_{mn}\}^2. \tag{3.2}$$

The subscript n on a quantity is to indicate that the quantity is to be evaluated under the assumption that the input to the system is unmodulated noise, while the subscript mn will be used to indicate that the quantity is to be evaluated under the assumption that the input is amplitude-modulated noise.

Under the assumption that the observations are independent, the difference $\Delta(T)$ of the observed outputs in the two cases,

$$\Delta(T) = w(T)_{mn} - w(T)_n, \tag{3.3}$$

will have mean value $\langle\Delta(T)\rangle = \langle w(T)\rangle_{mn} - \langle w(T)\rangle_n$ and variance equal to $\sigma_{mn}^2 + \sigma_n^2$. Using the statistical theory of testing hypotheses, if the probability

distribution of $\Delta(T)$ is known, by testing the hypothesis that $\langle \Delta(T) \rangle = 0$ one can detect the presence or the absence of signal. To any preassigned probability p one can find a number, K, such that

$$| \Delta(T) | \leq K[\sigma_{mn}^2 + \sigma_n^2]^{\frac{1}{2}} \tag{3.4}$$

with probability p. As the detection process, one then adopts the following rule. One decides that noise alone is present if the difference of the observed outputs lies in the range given by Eq. (3.4), and that signal (in our case, modulated noise) is present if the difference of the observed outputs lies outside the range given by Eq. (3.4).

Motivated by the foregoing considerations, we therefore adopt as our detection criterion, denoted by $D(T)$, the following ratio:

$$D(T) = \frac{\langle \overline{w}(T) \rangle_{mn} - \langle \overline{w}(T) \rangle_n}{[\sigma_{mn} + \sigma_n^2]^{\frac{1}{2}}} \tag{3.5}$$

It is presumed that the larger $D(T)$ is for a given detection system, the better will be the system's performance, and the more likely it will be to detect the presence of signal.

In many cases, it can be assumed that the fluctuation term for the case of signal and noise in the input is roughly equal to the fluctuation term σ_n for the case of noise alone in the input. We then have for the detection criterion

$$D(T) = \frac{\langle \overline{w}(T) \rangle_{mn} - \langle \overline{w}(T) \rangle_n}{\sqrt{2}\sigma_n} \tag{3.6}$$

In view of assumption (1.9), one could show that σ_{mn} is roughly equal to σ_n, and we therefore use Eq. (3.6) for the detection criterion.

We now turn our attention to the problem of expressing $D(T)$ in terms of the statistics of the noise and of the modulating function.

It is clear that

$$\langle \overline{w}(T) \rangle_n = T \langle v_1(t) v_2(t) \rangle_n \tag{3.7}$$

$$\langle \overline{w}(T) \rangle_{mn} = T \frac{1}{T} \int_0^T \langle v_1(t) v_2(t) \rangle_{mn} \, dt \tag{3.8}$$

Next, one may show [2] that

$$\langle \overline{w}^2(T) \rangle_n = 2 \int_0^T d\tau (T - \tau) \langle v_1(t) v_2(t) v_1(t + \tau) v_2(t + \tau) \rangle_n \tag{3.9}$$

Therefore

$$\sigma_n^2[\overline{w}(T)] = 2 \int_0^T d\tau (T - \tau) \{ \langle v_1(t) v_2(t) v_1(t + \tau) v_2(t + \tau) \rangle_n - \langle v_1(t) v_2(t) \rangle_n^2 \} \tag{3.10}$$

We may then obtain the following expression for the limit of $D(T)$ as the in-

tegration time T becomes infinite:

$$\lim_{T\to\infty}\frac{D^2(T)}{T}=\frac{\{[\langle v_1(t)v_2(t)\rangle_m]_{TAv}-\langle v_1(t)v_2(t)\rangle_n\}^2}{2\int_0^\infty d\tau\{\langle v_1(t)v_2(t)v_1(t+\tau)v_2(t+\tau)\rangle_n-\langle v_1(t)v_2(t)\rangle_n^2\}} \tag{3.11}$$

if we assume, as will turn out to be the case, that the integrand in the denominator of the right hand side of Eq. (3.11) is absolutely integrable.

In appendices B and C, we obtain expressions for the various terms which appear on the right hand side of Eq. (3.11), in terms of the power spectra $G_y(\omega)$, $G_s(\omega)$, and $G_g(\omega)$, and the filter transfer functions $H_1(\omega)$, $H_2(\omega)$, $K_1(\omega)$ and $K_2(\omega)$. Thus the problem of expressing $D(T)$ in terms of the statistics of the noise and the modulating function has been solved in principle.

4. Application to the Case of Gaussian Filters. In this section, we illustrate the foregoing results by applying them to the case where the noise spectra are flat, and the spectrum of the modulating function and the filter transfer functions have Gaussian shapes. While not physically realizable, Gaussian shaped filters are often good approximations, for mathematical purposes, of actual filters. We also consider only the case where $H_1(\omega)=H_2(\omega)=H(\omega)$, and $K_1(\omega)=K_2(\omega)=K(\omega)$.

We thus assume that, up to phase factors, the filter transfer functions $H(\omega)$ and $K(\omega)$ are given by Gaussian functions, as follows:

$$H(\omega)=\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega_H}{\sigma_H}\right)^2\right]+\exp\left[-\tfrac{1}{2}\left(\frac{\omega+\Omega_H}{\sigma_H}\right)^2\right] \tag{4.1}$$

$$K(\omega)=\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega_K}{\sigma_K}\right)^2\right]+\exp\left[-\tfrac{1}{2}\left(\frac{\omega+\Omega_K}{\sigma_K}\right)^2\right] \tag{4.2}$$

For the power spectrum $G_g(\omega)$ of the modulating function $g(t)$, we assume a Gaussian function centered at roughly the same frequency as is $K(\omega)$:

$$G_g(\omega)=P_g\frac{1}{\sqrt{2\pi}\sigma_g}\left\{\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega_K}{\sigma_g}\right)^2\right]+\exp\left[-\tfrac{1}{2}\left(\frac{\omega+\Omega_K}{\sigma_K}\right)^2\right]\right\} \tag{4.3}$$

There is no difficulty, however, in treating any other form of spectrum.

The noise power spectra $G_y(\omega)$ and $G_s(\omega)$ are assumed to be flat and identically equal to constants G_y and G_s respectively.

We finally assume that

$$\frac{\sigma_H^2}{\sigma_H^2+\sigma_K^2+\sigma_g^2}\simeq 1,\qquad \exp\left[-\left(\frac{\Omega_K}{\sigma_H}\right)^2\right]\simeq 1 \tag{4.4}$$

Physically, the meaning of condition (4.4) is that the $K(\omega)$ filters and the spectrum $G_g(\omega)$ of the modulating function are narrow band compared with the $H(\omega)$ filter, and also that the $K(\omega)$ filter is low frequency compared with the bandwidth of the $H(\omega)$ filter.

Under these assumptions, we obtain the following approximate expressions:

$$\frac{\langle \overline{w}(T)\rangle_{ms} - \langle \overline{w}(T)\rangle_n}{T} = 4\pi m^2 P_g G_y^2 \sigma_H^2 \left\{\left[1 + \left(\frac{\sigma_g}{\sigma_K}\sqrt{2}\right)^2\right]^{-\frac{1}{2}} + \frac{3}{\sqrt{2}}\frac{\sigma_K}{\sigma_H} + \frac{1}{\sqrt{2}}\frac{\sigma_K}{\sigma_H}\frac{G_z}{G_y}\right\} \tag{4.5}$$

$$\frac{\sigma_n^2[\overline{w}(T)]}{T} = (2\pi)^{\frac{1}{2}}(G_y + G_z)^4 \sigma_H^2 \sigma_K \left\{1 + (4\sqrt{2} + 2)\frac{\sigma_K}{\sigma_H}\right\} \tag{4.6}$$

As a measure of the ratio of the bandwidth of the spectrum of the modulating function to the bandwidth of the power transfer functions $|H(\omega)|^2$ and $|K(\omega)|^2$, define respectively

$$R_H = \frac{\sigma_g}{\sigma_H}\sqrt{2}, \qquad R_K = \frac{\sigma_g}{\sigma_K}\sqrt{2}. \tag{4.7}$$

In view of Eqs. (3.6), (4.5), and (4.6), we have the following approximate expression for the value of the detection criterion:

$$\frac{D^2(T)}{T} = \frac{1}{\sqrt{\pi}}(2m^2P_g)^2 \frac{\sigma_H^2}{\sigma_g} \frac{1}{R_K + (4\sqrt{2} + 2)R_H} \left\{R_K(1 + R_K^2)^{-1} + \frac{1}{\sqrt{2}} R_H\left(3 + \frac{G_z}{G_y}\right)\right\}^2 \tag{4.8}$$

In many cases, R_H may be considered small. Then Eq. (4.8) may be simplified by dropping terms involving R_H.

5. Conclusions. In the foregoing, we have developed formulae which enable us to study the behavior, and the design, of the detection system of Fig. 1. We have considered a detection criterion which may provide a possible general approach to the problem of designing good (though not necessarily optimum) detection systems [4].

We have also made two calculations which may be of general interest. We have computed in Appendices B and C, the cross-correlation $\langle v_1(t)v_2(t+\tau)\rangle_{TAv}$ of the inputs $v_1(t)$ and $v_2(t)$ of the multiplier for two kinds of original input to the system, stationary Gaussian noise and amplitude-modulated noise. We have also computed, in Appendix B, the autocorrelation $\langle w(t)w(t+\tau)\rangle$ of the output $w(t)$ of the multiplier when the input of the system is stationary Gaussian noise. From these correlation functions, the corresponding power spectra can be obtained by means of the Wiener-Khintchine relations. Some of the results given here may also be used to give an alternative analysis of a new type of interferometer which has been considered by Brown and Twiss [3].

Appendix A—The Higher Moments of Gaussian Noise. Let $y(t)$ be a stationary random time function whose probability distribution is Gaussian and whose mean value $\langle y(t)\rangle$ is zero for all t. Then all odd order moments of $y(t)$ vanish,

and the even order moments may be expressed in terms of the second order moments by the following formula. Let n be an even integer, and let $t_1, \cdots, t_n$ be points of time, some of which may coincide. Then (see [5], [6])

$$\langle y(t_1) \cdots y(t_n) \rangle = \sum \langle y(t_{i_1}) y(t_{i_2}) \rangle \cdots \langle y(t_{i_{n-1}}) y(t_{i_n}) \rangle \tag{A.1}$$

where the sum is taken over all possible ways of dividing the n points into $n/2$ combinations of pairs. The number of terms in the summation is equal to $1 \cdot 3 \cdots (n-3)(n-1)$.

As a special case of Eq. (A.1) we compute the fourth product moment of the squares of 4 Gaussian random variables u_1, u_2, u_3, and u_4. We have

$$\begin{aligned}
&\langle (u_1^2 - \langle u_1^2 \rangle)(u_2^2 - \langle u_2^2 \rangle)(u_3^2 - \langle u_3^2 \rangle)(u_4^2 - \langle u_4^2 \rangle) \rangle \\
&\quad = 4\langle u_1 u_2 \rangle^2 \langle u_3 u_4 \rangle^2 + 4\langle u_1 u_3 \rangle^2 \langle u_2 u_4 \rangle^2 + 4\langle u_1 u_4 \rangle^2 \langle u_2 u_3 \rangle^2 \\
&\qquad + 16\langle u_1 u_2 \rangle \langle u_1 u_4 \rangle \langle u_2 u_3 \rangle \langle u_3 u_4 \rangle + 16\langle u_1 u_2 \rangle \langle u_1 u_3 \rangle \langle u_2 u_4 \rangle \langle u_3 u_4 \rangle \\
&\qquad + 16\langle u_1 u_3 \rangle \langle u_1 u_4 \rangle \langle u_2 u_3 \rangle \langle u_2 u_4 \rangle
\end{aligned} \tag{A.2}$$

Appendix B—The Second and Fourth Moments of Filtered Squared Gaussian Noise. Let $u_1(t)$ and $u_2(t)$ be stationary Gaussian random functions, with auto- and cross-correlation functions

$$R_{11}(\tau) = \langle u_1(t) u_1(t+\tau) \rangle = \int_0^\infty \cos \omega\tau G_{11}(\omega)\, d\omega,$$

$$R_{22}(\tau) = \langle u_2(t) u_2(t+\tau) \rangle = \int_0^\infty \cos \omega\tau G_{22}(\omega)\, d\omega,$$

$$R_{12}(\tau) = \langle u_1(t) u_2(t+\tau) \rangle = \langle u_2(t) u_1(t+\tau) \rangle = \int_0^\infty \cos \omega\tau G_{12}(\omega)\, d\omega.$$

Let for $i = 1, 2$, $v_i(t)$ represent the output of a filter whose input is the square of $u_i(t)$, so that

$$v_1(t) = \int d\alpha k_i(\alpha) u_i^2(t - \alpha)$$

where the filter frequency transfer functions

$$K_i(\omega) = \int d\alpha k_1(\alpha) e^{i\alpha\omega}$$

pass no D.C., so that $K_i(0) = 0$. Therefore $\langle v_i(t) \rangle = 0$.

Then, for i and j equal to 1 or 2,

$$\langle v_i(t) v_j(t+\tau) \rangle = \tfrac{1}{2} \int d\omega K_i(\omega) K_j^*(\omega) e^{i\tau\omega} \int d\mu G_{ij}(\mu) G_{ij}(\omega - \mu)$$

From Eq. (A.2), it may be inferred that the fourth moment is given by

$$\langle v_1(t)v_2(t)v_1(t+\tau)v_2(t+\tau)\rangle$$

$$= \langle v_1(t)v_2(t)\rangle\langle v_1(t+\tau)v_2(t+\tau)\rangle + \langle v_1(t)v_1(t+\tau)\rangle\langle v_2(t)v_2(t+\tau)\rangle$$

$$+ \langle v_1(t)v_2(t+\tau)\rangle\langle v_2(t)v_1(t+\tau)\rangle$$

$$+ \int d\omega e^{i\tau\omega} \int d\lambda G_{11}(\omega-\lambda)G_{22}(\lambda) \left| \int d\mu G_{12}(\lambda-\mu)K_1(\omega-\mu)K_2^*(\mu) \right|^2$$

$$+ \int d\omega e^{i\tau\omega} \int d\lambda G_{12}(\omega-\lambda)G_{12}^*(\lambda) \int d\mu_1 G_{12}(\lambda-\mu_1)K_1(\omega-\mu_1)K_2^*(\mu_1)$$

$$\int d\mu_2 G_{12}(\lambda-\mu_2)K_1^*(\mu_2)K_2(\omega-\mu_2)$$

$$+ \int d\omega e^{i\tau\omega} \iiint d\lambda\, d\mu_1\, d\mu_2 K_1(\omega-\mu_1)K_2^*(\mu_1)K_1^*(\omega-\mu_2)K_2(\mu_2)$$

$$G_{12}(\lambda-\mu_1)G_{12}^*(\lambda-\mu_2)G_{11}(\omega+\lambda-\mu_1-\mu_2)G_{22}(\lambda)$$

We next obtain its integral.

$$\int_{-\infty}^{\infty} d\tau \{\langle v_1(t)v_2(t)v_1(t+\tau)v_2(t+\tau)\rangle - \langle v_1(t)v_2(t)\rangle^2\}$$

$$= \frac{\pi}{2}\int d\lambda \mid K_1(\lambda)K_2^*(\lambda)\mid^2 \int d\mu_1 G_{11}(\mu_1)G_{11}(\mu_1+\lambda) \int d\mu_2 G_{22}(\mu_2)G_{22}(\mu_2+\lambda)$$

$$+ \frac{\pi}{2}\int d\lambda \mid K_1(\lambda)K_2^*(\lambda)\mid^2 \left| \int d\mu G_{12}(\mu)G_{12}^*(\mu+\lambda) \right|^2$$

$$+ 2\pi \int d\lambda G_{11}(\lambda)G_{22}(\lambda) \left| \int d\mu K_1(\mu)K_2^*(\mu)G_{12}(\mu+\lambda) \right|^2$$

$$+ 2\pi \int d\lambda \mid G_{12}(\lambda)\mid^2 \left| \int d\mu K_1(\mu)K_2^*(\mu)G_{12}(\mu+\lambda) \right|^2$$

$$+ 2\pi \int d\lambda K_1^*(\lambda)K_2(\lambda) \iint d\mu_1\, d\mu_2 K_1(\lambda+\mu_1+\mu_2)K_2^*(\lambda+\mu_1+\mu_2)$$

$$G_{11}(\mu_1)G_{22}(\mu_2)G_{12}(\mu_1+\lambda)G_{12}(\mu_2+\lambda)$$

Appendix C—The Crosscorrelation of Squared Filtered Amplitude-Modulated Gaussian Noise in the Presence of Background Noise. We compute, in this appendix, the crosscorrelation $[\langle v_1(t)v_2(t+\tau)\rangle_{mn}]_{\tau,v}$ of the outputs of the filters K_1 and K_2 of the system of Fig. 1, when the input of the system is given by Eq. (1.6) as the sum of an amplitude-modulated Gaussian noise plus a background noise. By a direct calculation, it may be verified that, up to terms involving m,

$$[\langle u_1^2(t)u_2^2(t+\tau)\rangle_{mn}]_{TAv} = \iiiint d\alpha_1\, d\alpha_2\, d\beta_1\, d\beta_2 h_1(\alpha_1)h_1(\alpha_2)h_2(\beta_1)h_2(\beta_2)$$

$$\begin{aligned}
&[\{R_{y+s}(\alpha_1-\alpha_2)R_{y+s}(\beta_1-\beta_2)+R_{y+s}(\tau+\alpha_1-\beta_1)R_{y+s}(\tau+\alpha_2-\beta_2)\\
&\qquad+R_{y+s}(\tau+\alpha_1-\beta_2)R_{y+s}(\tau+\alpha_2-\beta_1)\}\\
&+m^2\{R_y(\alpha_1-\alpha_2)R_y(\beta_1-\beta_2)+R_y(\tau+\alpha_1-\beta_1)R_y(\tau+\alpha_2-\beta_2)\\
&\qquad+R_y(\tau+\alpha_1-\beta_2)R_y(\tau+\alpha_2-\beta_1)\}\\
&\quad\{R_g(\alpha_1-\alpha_2)+R_g(\beta_1-\beta_2)+R_g(\tau+\alpha_1-\beta_1)+R_g(\tau+\alpha_1-\beta_2)\\
&\qquad+R_g(\tau+\alpha_2-\beta_1)+R_g(\tau+\alpha_2-\beta_2)\}\\
&\qquad+m^2\{R_s(\alpha_1-\alpha_2)R_y(\beta_1-\beta_2)R_g(\beta_1-\beta_2)\\
&\qquad+R_s(\beta_1-\beta_2)R_y(\alpha_1-\alpha_2)R_g(\alpha_1-\alpha_2)\\
&\qquad+R_s(\tau+\alpha_1-\beta_1)R_y(\tau+\alpha_2-\beta_2)R_g(\tau+\alpha_2-\beta_2)\\
&\qquad+R_s(\tau+\alpha_2-\beta_2)R_y(\tau+\alpha_1-\beta_1)R_g(\tau+\alpha_1-\beta_1)\\
&\qquad+R_s(\tau+\alpha_1-\beta_2)R_y(\tau+\alpha_2-\beta_1)R_g(\tau+\alpha_2-\beta_1)\\
&\qquad+R_s(\tau+\alpha_2-\beta_1)R_y(\tau+\alpha_1-\beta_2)R_g(\tau+\alpha_1-\beta_2)\}]
\end{aligned}$$

where we define $R_{y+s}(\tau) = R_y(\tau) + R_s(\tau)$ and $G_{y+s}(\omega) = G_y(\omega) + G_s(\omega)$.

By writing these correlation functions in terms of their corresponding power spectra, it follows that

$$\begin{aligned}
&[\langle v_1(t)v_2(t+\tau)\rangle_{mn}]_{TAv} = \tfrac{1}{2}\int d\lambda e^{i\tau\lambda}K_1(\lambda)K_2^*(\lambda)\int d\mu G_{y+s}(\mu)G_{y+s}(\lambda-\mu)\\
&\quad H_1(\mu)H_2^*(\mu)H_1(\lambda-\mu)H_2^*(\lambda-\mu)\\
&\qquad+\tfrac{1}{2}m^2\int d\lambda e^{i\tau\lambda}K_1(\lambda)K_2^*(\lambda)G_g(\lambda)\int d\mu_1 G_y(\mu_1)H_1^*(\mu_1)H_1(\mu_1+\lambda)\\
&\qquad\qquad\int d\mu_2 G_y(\mu_2)H_2^*(\mu_2)H_2(\mu_2-\lambda)\\
&\qquad+\tfrac{1}{2}m^2\int d\lambda e^{i\tau\lambda}K_1(\lambda)K_2^*(\lambda)\int d\nu G_g(\nu)\int d\mu G_y(\mu)G_y(\lambda-\mu)\\
&\quad H_1(\mu)H_1(\lambda-\mu)H_2^*(\mu-\nu)H_2^*(\lambda+\nu-\mu)\\
&\qquad+\tfrac{1}{2}m^2\int d\lambda e^{i\tau\lambda}K_1(\lambda)K_2^*(\lambda)\int d\nu G_g(\nu)\int d\mu G_y(\mu-\nu)G_y(\mu-\lambda)\\
&\quad H_1(\mu)H_1^*(\mu-\lambda)\{H_2^*(\mu)H_2(\mu-\lambda)+H_2^*(\mu-\nu)H_2(\mu-\nu-\lambda)\}\\
&\qquad+\tfrac{1}{2}m^2\int d\lambda e^{i\tau\lambda}K_1(\lambda)K_2^*(\lambda)\int d\nu G_g(\nu)\int d\mu G_y(\mu-\nu)G_s(\mu-\lambda)\\
&\qquad\qquad H_1(\mu)H_2^*(\mu)H_1^*(\mu-\lambda)H_2(\mu-\lambda)
\end{aligned}$$

Appendix D—The Product and Integral of Gaussian Functions. It is often computationally convenient to assume Gaussian shapes for filter transfer functions and signal and noise spectra, in view of the following two formulas for the product and integral of Gaussian functions:

$$\exp\left[-\left(\frac{\omega-\Omega_1}{\sigma_1}\right)^2\right]\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega_2}{\sigma_2}\right)^2\right]$$
$$=\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega}{\sigma}\right)^2\right]\exp\left[-\tfrac{1}{2}\frac{(\Omega_1-\Omega_2)^2}{\sigma_1^2+\sigma_2^2}\right]$$

where

$$\Omega=\frac{\Omega_1\sigma_2^2+\Omega_2\sigma_1^2}{\sigma_1^2+\sigma^2},\qquad \sigma^2=\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

and

$$\int_{-\infty}^{\infty}d\omega\,\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega_1}{\sigma_1}\right)^2\right]\exp\left[-\tfrac{1}{2}\left(\frac{\omega-\Omega_2}{\sigma_2}\right)^2\right]$$
$$=\sqrt{2\pi}\left[\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right]^{\frac{1}{2}}\exp\left[-\tfrac{1}{2}\frac{(\Omega_1-\Omega_2)^2}{\sigma_1^2+\sigma_2^2}\right]$$

REFERENCES

[1] Deutsch, R. E., "Detection of Modulated Noise-Like Signals," *I.R.E. Professional Group on Information Theory Transactions*, No. 3 (1954), pp. 106–123.

[2] Davenport, W. B., Johnson, R. A., and Middleton, D., "Statistical Errors in Measurements on Random Time Functions," *Journal of Applied Physics*, Vol. 23 (1952) p. 379, Eq. (2.5).

[3] Brown, R. H., and Twiss, R. Q., "A New Type of Interferometer for Use in Radio Astronomy," *Phil. Mag.*, Vol. 45 (1954) pp. 663–682.

[4] Parzen, E., "Detection and Hypothesis Testing by Linear Methods" (abstract), *Annals of Mathematical Statistics*, Vol. 26 (1955), p. 159.

[5] Wang, M. C. and Uhlenbeck, G. E., "On the Theory of the Brownian Motion II," *Reviews of Modern Physics*, Vol. 17 (1949), p. 332, Eq. (4.2).

[6] Lawson, J. L., and Uhlenbeck, G. E., *Threshold Signals* (McGraw-Hill Book Co., New York, 1950), p. 68, Eq. (12).

Columbia University
Hudson Laboratories
Dobbs Ferry, N. Y.

(Received January 7, 1955, Revised April 23, 1956)

ON THE RATIO OF CONSECUTIVE EIGENVALUES[1]

By L. E. Payne, G. Pólya and H. F. Weinberger

Introduction. 1. We consider a domain D in the x, y-plane, its boundary C, and three classical eigenvalue problems, concerned with the membrane, the clamped plate and the buckling of a plate, respectively:

$$\Delta u + \lambda u = 0 \text{ in } D, \quad u = 0 \text{ on } C, \tag{1}$$

$$\Delta\Delta u - \mu u = 0 \text{ in } D, \quad u = \partial u/\partial n = 0 \text{ on } C, \tag{2}$$

$$\Delta\Delta u + \nu\Delta u = 0 \text{ in } D, \quad u = \partial u/\partial n = 0 \text{ on } C. \tag{3}$$

As usual, Δ denotes the Laplace operator and n the normal to C. We use the notation

$$\lambda_1, \lambda_2, \lambda_3, \cdots$$

$$\mu_1, \mu_2, \mu_3, \cdots$$

$$\nu_1, \nu_2, \nu_3, \cdots$$

for the successive eigenvalues in problems (1), (2), and (3), respectively; $\lambda_n \leqq \lambda_{n+1}$, $\mu_n \leqq \mu_{n+1}$, $\nu_n \leqq \nu_{n+1}$ for $n \geqq 1$.

We shall prove that, *independently of n and of the shape of D*,

$$\lambda_{n+1} \leqq 3\lambda_n, \tag{4}$$

$$\mu_{n+1} \leqq 9\mu_n. \tag{5}$$

In fact, we shall prove a little more, namely

$$\lambda_{n+1} \leqq \lambda_n + \frac{2(\lambda_1 + \lambda_2 + \cdots + \lambda_n)}{n}, \tag{4'}$$

$$\mu_{n+1} \leqq \mu_n + \frac{8(\mu_1 + \mu_2 + \cdots + \mu_n)}{n}. \tag{5'}$$

The proofs for (4') and (5') start out along the same line and diverge only slightly from the point of separation onwards. Unfortunately, the method that works equally well for the eigenvalue problems (1) and (2) runs into additional complications when it is applied to (3). We prove here only that

$$\nu_2 < 3\nu_1 \tag{6}$$

and leave aside the higher eigenvalues ν_n. We shall also sketch a proof for

$$\lambda_2 + \lambda_3 \leqq 6\lambda_1. \tag{7}$$

[1] This paper was prepared partly under the sponsorship of the Office of Naval Research, and partly under the sponsorship of the United States Air Force Office of Scientific Research, Air Research and Development Command. The results were briefly announced in the *Comptes Rendus*, Paris, Vol. 241, 1955, pp. 917-919.

We have proved relations analogous to (7) also for the problems (2) and (3), but the constants obtained seem to be less good and so we leave them aside too.

For the circular membrane $\lambda_2/\lambda_1 = 2.539$ approximately; perhaps inequality (4) remains generally true if we substitute in it this value for the bound 3 that we have succeeded in proving. At any rate, there are some indications in favor of this conjecture.

2. This paper consists of five parts. Part I applies equally to the case of the membrane and that of the clamped plate and prepares the ground for the proof of (4′) and (5′). Parts II, III, and IV prove (4′), (5′), and (6), respectively. Part V contains additional remarks.

As double integrals extended over the domain D occur rather frequently, we shall use throughout the paper the abbreviation $\int f$ for the full symbol $\int_D\int f(x, y)\, dx\, dy$.

If the limits of a summation are not indicated, the subscript ranges from 1 to n. If just one subscript appears under the sign $\sum$, the summation is extended to that subscript. For example, $\sum u_i^2$ must be understood as an abbreviation of the full symbol $\sum_{i=1}^{n} u_i^2$.

The letters x and y used as subscripts denote partial derivatives in the usual way; we use i, j, or k as subscripts for summation, never x or y. For example, u_{ix} stands for $\partial u_i/\partial x$.

I. Membrane and Plate. 3. We regard as known the first n eigenfunctions

$$u_1, u_2, \cdots u_n$$

and consider such properties which belong equally to problems (1) and (2). The principal properties of this kind are

$$u_i = 0 \quad \text{on} \quad C, \tag{1.1}$$

$$\int u_i u_j = \begin{cases} 1 & \text{when} \quad i = j, \\ 0 & \text{when} \quad i \neq j. \end{cases} \tag{1.2}$$

For reasons which will be apparent later, it is desirable to choose the direction of the coordinate axes so that

$$\sum \int u_{ix}^2 = \sum \int u_{iy}^2. \tag{1.3}$$

In fact, this choice is possible. Assume that in a certain position of the coordinate axes, the left hand side of (1.3) is greater than the right hand side. A rotation of the coordinate system about the origin through 90° interchanges the axes and reverses the inequality. Therefore, there is an intermediate position in which (1.3) holds. (The admissibility of (1.3) could be also proved without appeal to continuity, by the explicit formulas of coordinate transformation.)

4. To obtain information about the next eigenvalue, we shall need an appropriate "trial-function." In fact, we shall use n trial-functions

$$\varphi_1 , \varphi_2 , \cdots \varphi_n .$$

Any one of these trial-functions, say φ_i, must satisfy two conditions:

$$\varphi_i = 0 \quad \text{on} \quad C, \tag{1.4}$$

$$\int \varphi_i u_j = 0 \tag{1.5}$$

for $i, j = 1, 2, \cdots n$. We choose

$$\varphi_i = xu_i - \textstyle\sum_k a_{ik} u_k \tag{1.6}$$

where a_{ik} are suitable constants. By virtue of (1.1), the condition (1.4) is satisfied. By virtue of (1.2), the condition (1.5) is equivalent to

$$a_{ij} = \int x u_i u_j = a_{ji} . \tag{1.7}$$

Therefore, we define a_{ij} by (1.7) and have (1.5). Observe that from (1.5) and (1.6) it follows that

$$\int \varphi_i^2 = \int x u_i \varphi_i . \tag{1.8}$$

5. We shall need later the sum

$$-2 \sum \int u_{ix} \varphi_i = -\sum \int x \, 2u_i u_{ix} + 2 \sum_i \sum_k a_{ik} \int u_{ix} u_k ; \tag{1.9}$$

we used (1.6). We integrate by parts, using (1.1). On the one hand

$$\int u_{ix} u_k = -\int u_{kx} u_i ;$$

from this "antisymmetry" and from the symmetry of a_{ik}, see (1.7), it follows that the double sum on the right hand side of (1.9) vanishes. On the other hand

$$-\int x 2u_i \, u_{ix} = \int u_i^2 = 1;$$

we used (1.2). Applying both remarks to (1.9), we obtain

$$-2 \sum \int u_{ix} \varphi_i = n. \tag{1.10}$$

We derive hence, by applying Schwarz's inequality and (1.3), that

$$n^2 \leqq 4 \sum \int \varphi_i^2 \sum \int u_{ix}^2 = 2 \sum \int \varphi_i^2 \sum \int u_{ix}^2 + u_{iy}^2 . \tag{1.11}$$

II. Membrane. 6. We regard as known the first n eigenfunctions

$$u_1 , u_2 , \cdots u_n$$

and the corresponding eigenvalues

$$\lambda_1 , \lambda_2 , \cdots \lambda_n$$

of problem (1). We add a property specific to this problem to the properties considered in Part I:

$$\Delta u_i + \lambda_i u_i = 0 \quad \text{in} \quad D. \tag{2.1}$$

From this, (1.1), and (1.2) it follows that

$$\int u_{ix}^2 + u_{iy}^2 = \lambda_i . \tag{2.2}$$

The next eigenvalue λ_{n+1} is characterized by the inequality

$$\lambda_{n+1} \leqq -\frac{\int \varphi_i \Delta \varphi_i}{\int \varphi_i^2} \tag{2.3}$$

valid for any function φ_i satisfying (1.4) and (1.5). Such a function is (1.6) for which we find, using (2.1), that

$$-\Delta \varphi_i = \lambda_i x u_i - 2u_{ix} - \textstyle\sum_k a_{ik} \lambda_k u_k .$$

We derive hence, using (1.8) and (1.5), that

$$-\int \varphi_i \Delta \varphi_i = \lambda_i \int \varphi_i^2 - 2 \int u_{ix} \varphi_i$$

and so (2.3) yields

$$\lambda_{n+1} \leqq \lambda_i - \frac{2 \int u_{ix} \varphi_i}{\int \varphi_i^2} \leqq \lambda_n - \frac{2 \int u_{ix} \varphi_i}{\int \varphi_i^2} .$$

Hence we obtain, using (1.10), (1.11), and (2.2) in succession

$$\lambda_{n+1} - \lambda_n \leqq -\frac{2 \sum \int u_{ix} \varphi_i}{\sum \int \varphi_i^2} = \frac{n}{\sum \int \varphi_i^2} \leqq \frac{2 \sum \int u_{ix}^2 + u_{iy}^2}{n} = \frac{2 \sum \lambda_i}{n}$$

which is the desired (4′).

III. Plate. 7. We regard now as known the first n eigenfunctions $u_1 , u_2 , \cdots u_n$ and the corresponding eigenvalues $\mu_1 , \mu_2 , \cdots \mu_n$ of problem (2). We add two properties specific to this problem to the properties considered in Part I:

$$\Delta \Delta u_i - \mu_i u_i = 0 \quad \text{in} \quad D, \tag{3.1}$$

$$\frac{\partial u_i}{\partial n} = 0 \quad \text{on} \quad C. \tag{3.2}$$

From (1.1), (1.2), (3.1), and (3.2) it follows that

$$\int (\Delta u_i)^2 = \mu_i \,. \tag{3.3}$$

The next eigenvalue μ_{n+1} is characterized by the inequality

$$\mu_{n+1} \leqq \frac{\int \varphi_i \Delta\Delta \varphi_i}{\int \varphi_i^2} \tag{3.4}$$

valid for any function φ_i satisfying (1.4), (1.5), and

$$\frac{\partial \varphi_i}{\partial n} = 0 \quad \text{on} \quad C. \tag{3.5}$$

Such a function is (1.6) ((3.5) is satisfied by virtue of (1.1) and (3.2)) for which we find, using (3.1),

$$\Delta\Delta \varphi_i = \mu_i x u_i + 4\Delta u_{ix} - \textstyle\sum_k a_{ik}\mu_k u_k \,.$$

We derive hence, using (1.8) and (1.5), that

$$\int \varphi_i \Delta\Delta \varphi_i = \mu_i \int \varphi_i^2 + 4 \int \varphi_i \Delta u_{ix}$$

and so (3.4) yields

$$\mu_{n+1} \leqq \mu_i + \frac{4 \cdot \int \varphi_i \Delta u_{ix}}{\int \varphi_i^2} \leqq \mu_n + \frac{4 \int \varphi_i \Delta u_{ix}}{\int \varphi_i^2} \,.$$

Hence we obtain

$$\mu_{n+1} - \mu_n \leqq \frac{4 \sum \int \varphi_i \Delta u_{ix}}{\sum \int \varphi_i^2} \,. \tag{3.6}$$

8. We consider the numerator of the right hand side of (3.6). It is analogous to the sum (1.9) and we can treat it similary. By (1.6)

$$2 \sum \int \varphi_i \Delta u_{ix} = \sum 2 \int x u_i \Delta u_{ix} - 2 \textstyle\sum_i \sum_k a_{ik} \int u_k \Delta u_{ix} \,. \tag{3.7}$$

We integrate by parts, using (1.1) and (3.2). On the one hand

$$\int u_k \Delta u_{ix} = \int u_k \frac{\partial \Delta u_i}{\partial x} = -\int u_{kx} \Delta u_i = -\int u_i \Delta u_{kx} \,;$$

from this "antisymmetry" and the symmetry of a_{ik}, see (1.7), it follows that

the double sum on the right hand side of (3.7) vanishes. On the other hand,

$$2\int x u_i \Delta u_{ix} = -2\int (x u_{ix} + u_i) u_{ixx} + x u_{iy} u_{ixy}$$

$$= -\int x(2u_{ix}u_{ixx} + 2u_{iy}u_{iyx}) - 2\int u_i u_{ixx}$$

$$= 3\int u_{ix}^2 + \int u_{iy}^2 ;$$

we used that $u_{ix} = u_{iy} = 0$ along C, which is implied by (1.1) and (3.2). Applying both remarks to (3.7), and using (1.3) in the sequel, we obtain

$$2\sum\int \varphi_i \Delta u_{ix} = 3\sum\int u_{ix}^2 + \sum\int u_{iy}^2 = 2\sum\int u_{ix}^2 + u_{iy}^2 .$$

Using this and (1.1), we can rewrite (3.6) and apply afterwards, in succession, (1.11), Schwarz's inequality, (1.2), and (3.3):

$$\mu_{n+1} - \mu_n \leqq \frac{4\sum\int u_{ix}^2 + u_{iy}^2}{\sum\int \varphi_i^2} = -\frac{4\sum\int u_i \Delta u_i}{\sum\int \varphi_i^2} = \frac{8\left(\sum\int u_i \Delta u_i\right)^2}{2\sum\int \varphi_i^2 \sum\int u_{ix}^2 + u_{iy}^2}$$

$$\leqq \frac{8\left(\sum\int u_i \Delta u_i\right)^2}{n^2} \leqq \frac{8\sum\int u_i^2 \sum\int (\Delta u_i)^2}{n^2} = \frac{8\sum \mu_i}{n}$$

which is the desired (5′).

IV. Buckling. 9. We assume as known the first eigenvalue ν_1 and the corresponding eigenfunction u of problem (3) so that

$$\Delta\Delta u + \nu_1 \Delta u = 0 \quad \text{in} \quad D, \qquad u = \frac{\partial u}{\partial n} = 0 \quad \text{on} \quad C. \tag{4.1}$$

We introduce the usual abbreviations

$$\int f_x g_x + f_y g_y = D[f, g], \qquad D[f, f] = D[f]. \tag{4.2}$$

It follows from (4.1) that

$$\int u\Delta\Delta u = \int (\Delta u)^2 = D[u_x] + D[u_y] = \nu_1 D[u]. \tag{4.3}$$

It is well known that

$$\nu_2 \leqq \frac{\int \varphi\Delta\Delta\varphi}{D[\varphi]} \tag{4.4}$$

provided that the trial function φ satisfies two conditions:

$$\varphi = \frac{\partial \varphi}{\partial n} = 0 \quad \text{on} \quad C, \tag{4.5}$$

$$D[u, \varphi] = 0. \tag{4.6}$$

10. Cover the domain D with matter of surface density $u_x^2 + u_y^2$ and suppose that the center of gravity of this mass distribution lies at the origin:

$$\int x(u_x^2 + u_y^2)\, dx\, dy = \int y(u_x^2 + u_y^2)\, dx\, dy = 0. \tag{4.7}$$

Now, we choose as trial function

$$\varphi = xu. \tag{4.8}$$

Thanks to (4.1), the condition (4.5) is satisfied by this choice. On the other hand

$$\int uu_x = 0, \tag{4.9}$$

see the boundary condition in (4.1), and

$$D[u, \varphi] = \int u_x(xu_x + u) + u_y x u_y = 0$$

by (4.8), (4.9), and (4.7). Thus condition (4.6) is also satisfied: (4.8) is an admissible trial function.

11. We have to compute the right hand side of (4.4) using (4.8). Since, by (4.1),

$$\int x 2uu_x + u^2 = 0 \tag{4.10}$$

we obtain for the denominator

$$D[\varphi] = \int (xu_x + u)^2 + (xu_y)^2 = \int x^2(u_x^2 + u_y^2). \tag{4.11}$$

We examine the numerator:

$$\int \varphi \Delta\Delta\varphi = -\nu_1 \int x^2 u \Delta u + 4 \int \varphi \Delta u_x. \tag{4.12}$$

Now

$$\begin{aligned} -\int x^2 u \Delta u &= \int (x^2 u_x + 2xu)u_x + x^2 u_y u_y \\ &= D[\varphi] - \int u^2 < D[\varphi]; \end{aligned} \tag{4.13}$$

see (4.1), (4.10), and (4.11). On the other hand, in view of (4.5), (4.1), and (4.8)

$$2\int \varphi\Delta u_x = -2D[\varphi, u_x] = -2\int (xu_x + u)u_{xx} + xu_y u_{xy}$$

(4.14)

$$= -\int x(2u_x u_{xx} + 2u_y u_{yx}) - 2\int uu_{xx} = 3\int u_x^2 + \int u_y^2.$$

12. From (4.4), (4.12), (4.13), and (4.14)

$$\nu_2 < \nu_1 + \frac{4\int \varphi\Delta u_x}{D[\varphi]} = \nu_1 + \frac{8\left(\int \varphi\Delta u_x\right)^2}{D[\varphi]\left(3\int u_x^2 + \int u_y^2\right)}. \tag{4.15}$$

Now, by the Schwarz inequality,

$$\left(\int \varphi\Delta u_x\right)^2 = (D[\varphi, u_x])^2 \leqq D[\varphi]D[u_x]. \tag{4.16}$$

The combination of (4.15) and (4.16) yields

$$\nu_2 < \nu_1 + \frac{8D[u_x]}{3\int u_x^2 + \int u_y^2}. \tag{4.17}$$

Yet, in view of (4.7), we could have chosen in Section 10 yu as trial function instead of xu and so we also have

$$\nu_2 < \nu_1 + \frac{8D[u_y]}{\int u_x^2 + 3\int u_y^2}. \tag{4.18}$$

Combining (4.17) with (4.18) and using (4.3), we obtain

$$\nu_2 < \nu_1 + \frac{8(D[u_x] + D[u_y])}{4D[u]} = 3\nu_1$$

which is the desired (6).

V. Remarks. 13. In proving (7), we emphasize only the main steps and leave out a few intermediate steps (which all occurred essentially in the foregoing, even under more complicated circumstances).

We assume as known the first eigenvalue λ_1 and the corresponding eigenfunction u of problem (1) ($u = u_1$ in the notation of Parts I and II). Therefore,

$$\Delta u + \lambda_1 u = 0 \quad \text{in} \quad D, \quad u = 0 \quad \text{on} \quad C, \quad \int u^2 = 1 \tag{5.1}$$

from which properties it follows that

$$\int u_x^2 + u_y^2 = \lambda_1. \tag{5.2}$$

Cover the domain D with matter of surface density u^2 and choose the coordinate system so that the center of gravity of this mass distribution coincides with the origin and the principal axes of inertia coincide with the coordinate axes:

$$\int xu^2 = \int yu^2 = \int xyu^2 = 0. \tag{5.3}$$

We wish to use a theorem of Poincaré[3] (which we did not need in the foregoing) and so we have to consider a trial-function that depends linearly on three parameters ξ, η, and ζ:

$$\varphi = \xi\, aux + \eta\, buy + \zeta u; \tag{5.4}$$

the positive constants a and b are so determined that

$$a^2 \int u^2x^2 = b^2 \int u^2y^2 = 1. \tag{5.5}$$

We find without difficulty that

$$\int \varphi^2 = \xi^2 + \eta^2 + \zeta^2, \tag{5.6}$$

$$-\int \varphi\Delta\varphi = (\lambda_1 + a^2)\xi^2 + (\lambda_1 + b^2)\eta^2 + \lambda_1\zeta^2. \tag{5.7}$$

By Poincaré's theorem, the first three eigenvalues of the membrane are not greater than the corresponding characteristic roots of the pair (5.6), (5.7) of quadratic forms, and so

$$\lambda_1 + \lambda_2 + \lambda_3 \leqq 3\lambda_1 + a^2 + b^2. \tag{5.8}$$

Now (cf. 4.10)

$$1 = \left(\int u^2\right)^2 = \left(-\int x2uu_x\right)^2 \leqq 4\int u^2x^2 \int u_x^2$$

(we have used Schwarz's inequality) and so, by (5.5), we have

$$a^2 \leqq 4\int u_x^2.$$

Similarly,

$$b^2 \leqq 4\int u_y^2$$

and so, by (5.2),

$$a^2 + b^2 \leqq 4\int u_x^2 + u_y^2 = 4\lambda_1$$

which, with (5.8), yields the desired (7).

[3] cf. G. Pólya and M. Schiffer, *Journal d'Analyse Mathematique*, Vol. 3, 1953, pp. 245–345 especially pp. 275–279.

14. Since $\lambda_1 \leqq \lambda_2 \leqq \lambda_3$, it follows from (7) that

$$\lambda_2 \leqq 3\lambda_1, \qquad \lambda_3 \leqq 5\lambda_1. \tag{5.9}$$

The first of these inequalities is contained in (4) which, however, would yield only $\lambda_3 \leqq 9\lambda_1$ instead of the sharper second inequality under (5.9). Even (4′) yields only $\lambda_3 \leqq 7\lambda_1$.

We have also proved results analogous to (5.9) for the first three eigenvalues of the clamped plate and the buckling problem (problems (2) and (3), respectively), yet instead of the 6 arising in (7) we found greater and, probably, less sharp numerical constants. Let us emphasize a consequence: also ν_3/ν_2 has a numerical bound independent of the shape of the domain D.

15. The method that, in Parts I, II and III, works satisfactorily for problems (1) and (2), hits a snag when applied to problem (3). The difficulty arises with the analogue of the matrix a_{ij}, see (1.6), (1.7). This matrix, with elements $\int xu_iu_j$, is symmetric, whereas our choice (1.6) of a trial function leads in problem (3) to the matrix with elements $D[u_i, xu_j]$, which is not symmetric.

16. There are indications that, of all membranes, the circular membrane attains the maximum value of the ratio $(\lambda_2 + \lambda_3)/\lambda_1$. There are also indications that, if we let vary both n and the shape of the membrane, λ_{n+1}/λ_n attains its maximum, approximately 2.539, when $n = 1$ and the membrane is circular. We list some of the indications.

(a) In surveying those few cases in which the eigenvalue problem of the membrane has been completely solved (circular sectors, rectangles, three particular triangles) we found no numerical values contradicting the conjectures stated. Yet there is subtler evidence too: if a figure A is "closer" to the circle than another figure B, figure A turns out to have a higher value both for λ_2/λ_1 and for $(\lambda_2 + \lambda_3)/\lambda_1$ than B.

We regard here the equilateral triangle closer to the circle than any other triangle, the square closer than any other quadrilateral, and so on.

(b) The first variation of $\lambda_2 + \lambda_2$, computed for the circle from the gradients of the corresponding eigenfunctions, turns out to be proportional to the first variation of λ_1 analogously computed. Such indication, however, is not too strong: it would equally suggest that λ_6/λ_1 is either a minimum or a maximum for the circle, yet, in fact, neither is the case.

(c) Professor Szegö to whom we communicated our conjecture concerning $(\lambda_2 + \lambda_3)/\lambda_1$ found it in agreement with his results concerning nearly circular membranes.[3]

(d) For a given shape, λ_{n+1}/λ_n need not steadily decrease as n increases: it is enough to refer here to the example of the circle. Yet λ_{n+1}/λ_n seems to decrease "on the whole" in some yet unclarified sense and so it is not unreasonable to expect that the maximum of this ratio is attained for $n = 1$.

UNIVERSITY OF MARYLAND AND STANFORD UNIVERSITY

(Received December 13, 1955)

[3] *Journal of Rational Mechanics and Analysis*, Vol. 3, 1954, pp. 343–356.

THE TRUNCATION ERROR IN A SEMI-DISCRETE ANALOG OF THE HEAT EQUATION*

By Norman E. Friedmann

1. Introduction. One of the most widely used methods for obtaining numerical solutions of heat problems, when simple forms of solution are unavailable, consists of replacing the differential operator by a difference analog. Questions of stability, and convergence associated with solutions of both explicit and implicit analogs have been extensively considered in many papers, (1)–(8).

It is known that upon passing from the discrete to continuous in time in the simple 4 point explicit approximation, that the resulting difference-differential (semi-discrete) analog is stable for all spatial lattice dimensions, (9). This characteristic has made possible the application of semi-discrete analogs to the solution of a general class of heat problems by means of continuous computers, (10)–(12).

In this paper we consider the problem of obtaining an estimate of the difference between an exact solution of the heat equation in a finite solid and an exact solution of a semi-discrete analog replacing the same problem. This difference is associated with the question of convergence of the difference solution to the differential solution, and is sometimes inappropriately called "lumping error". The usage presumably stems from the lumped circuit versus distributed electrical circuit concept of the heat problem (See 12). In keeping with the terminology currently used when referring to discrete analogs, we shall describe the difference as the *truncation error*.

In obtaining expressions for the error, a number of options exist: An exact evaluation or estimate of *local error*, which is a function of both time and position, or an exact evaluation or estimate of *global error* which is a function of time alone. Depending on the particular application in mind, one or more of these approaches may be preferred over the remainder.

In their studies Klein, Touloukian and Eaton (13) obtained exact local error expressions for the problem under study here:

$$\mu_{xx} = (1/\alpha)\mu_t\,; \qquad \mu(0, t) = F, \qquad \mu(l, t) = \mu(x, 0) = 0$$

where F is a constant. Their results were expressed in terms of a Fourier number at the mid-points of each mesh interval. An extension of their analysis to include the treatment of time varying boundary conditions is not discussed, however, a superposition principle and a mean value theorem could be applied to obviate this shortcoming. In addition, bounds on the error at small Fourier numbers are not established. They remark, in this respect that "The high positive departures during the early part of the transient may in some cases be of little concern, as errors of hundreds of percent may actually be equivalent to only a small percentage of the applied potential (temperature)".

* The author is greatly indebted to Dr. Myron Tribus of the University of California, Los Angeles, for his numerous suggestions and criticism during the writing of this paper. Thanks are also due to Ed DeLand who critically examined the manuscript.

Lawson and McGuire (14) treat a problem, comparable to that undertaken by Klein, et al (23), for the semi-infinite solid. They use methods similar to those found in (13) and consider only exact evaluation of local error.

Using methods similar to those found in (13) and (14), Clarke (15), gives exact equations for local error for the heat equation under the following boundary and initial conditions:

$$\mu_x(0, t) = 0, \qquad \mu_x(l, t) = q/k; \qquad \mu(x, 0) = 0.$$

His analysis is confined to the plane $x = 0$ although this is not an inherent limitation of the method used.

Paschkis and Heisler (16), in a study of a general class of errors arising in the use of continuous computers, have departed from the procedure employed by many of the writers. Taking a predetermined mesh size (number) as a control they have experimentally compared results for other larger mesh sizes. Utilizing this approach they study the problem examined by Klein, et al.

Finally in reference (10) an entirely different concept of the error problem is introduced. The problem

$$\mu_{xx} = (1/\alpha)\mu_t\,; \qquad \mu(x, 0) = \mu_0\,, \qquad \mu(0, t) = \mu(l, t) = 0$$

is considered. The percentage deviation in the eigenvalues for each mode of the separated equation is then evaluated in terms of mesh size using the exact eigenvalue equation for comparison. The method permits only the estimate of error in time, not spatial position. A separate analysis for the latter would be necessary.

Thus, in all results obtained to date, attention has been focussed on an exact determination of local error. From the point of view of analog computation, it is more desirable to be able to make advance predictions of the maximum error, *global error*, than it is to know the exact *local error*. While exact global error expressions are desirable, our results are a compromise in that an upper bound on global error is obtained. Having performed computations on the basis of these error bounds, then one may wish to make an exact evaluation of local error if this is possible.

This upper bound on global error is expressed in terms of such easily determined quantities as mesh size and Fourier number. The results apply to a finite solid with prescribed (constant) boundary and initial temperatures, and may supply an insight into anticipated error under other boundary conditions for other geometrical configurations.

2. The Problem. Let the solution domain D be bounded by $0 < x \leqq l$, and open in time, $t > 0$. The prescribed continuous boundary function is assumed to take on the value F at $x = 0$ and vanish at $x = l$. We denote the Laplacian operator by ∇^2 and define $\nabla_h^2 w$ by

$$\nabla_h^2 w = h^{-2}[w(x + h, t) + w(x - h, t) - 2w(x, t)]$$

and $\nabla_x^2 w$ by

$$\nabla_x^2 w = \partial^2 w/\partial x^2.$$

For any value of mesh spacing h such that $N = lh^{-1}$ is an integer, we denote by $\mu_h(t)$ the function, defined at all netpoints (nh, t) in D, where n is an integer, and including the boundaries for which

$$\nabla_h^2 \mu_h = \frac{1}{\alpha} \frac{\partial \mu_h}{\partial t} \tag{2.1}$$

and

$$\mu_0(t) = F, \qquad \mu_N(t) = 0 \tag{2.2}$$

while initially

$$\mu_h(0) = 0. \tag{2.3}$$

The function which satisfies the problem consisting of Eq. (2.1)–(2.3) is

$$\mu_h(t) = \sum_{k=1}^{N} C_k\left(k, \frac{k\pi}{N}\right) g\left(t, \frac{k\pi}{2N}\right) \sin\left(\frac{k\pi x}{l}\right) \qquad 0 < x \leqq l \tag{2.4}$$

where $C_k(k, y)$ is the Fourier interpolation coefficient given by

$$C_k(k, y) = 2\alpha k l^{-2} \pi \left(\frac{\sin y}{y}\right) \tag{2.5}$$

and

$$g(t, y) = F \int_0^t \exp - \left[\alpha k^2 \pi^2 l^{-2} \left(\frac{\sin y}{y}\right)^2 (t - \tau)\right] d\tau. \tag{2.6}$$

The result (2.4) follows from classical methods and can be easily verified by substitution into Eq. (2.1) and using the sum,

$$\frac{1}{N} \sum_{k=1}^{N} \frac{\sin (n/N)k\pi}{\tan k\pi/2N} = 1 - \left(\frac{n}{N}\right); \qquad 0 < n \leqq N.$$

As mentioned in (17), the solution of

$$\nabla_x^2 \mu = \frac{1}{\alpha} \frac{\partial \mu}{\partial t} \tag{2.7}$$

which satisfies the boundary conditions

$$\mu(0, t) = F, \qquad \mu(l, t) = 0 \tag{2.8}$$

and the initial condition

$$\mu(x, 0) = 0 \tag{2.9}$$

is

$$\mu(x, t) = \sum_{k=1}^{\infty} C_k(k, 0) g(t, 0) \sin\left(\frac{k\pi x}{l}\right). \qquad 0 < x \leqq l. \tag{2.10}$$

At a netpoint (nh, t) the exact solution (2.4) of (2.1) differs from the exact solution (2.10) of (2.7) by the truncation error $|\mu - \mu_h|$. Constructing this difference for the present case we write using (2.4) and (2.10)

$$|\mu - \mu_h| \leqq E_1 + E_2 + E_3 \tag{2.11}$$

where

$$E_1 \leqq \frac{2F}{\pi} \sum_{k=1}^{N} \left| \gamma_k\left(k, \frac{k\pi}{2N}\right) - \gamma_k(k, 0) \right| v(t, 0) \tag{2.12}$$

$$E_2 \leqq \frac{2F}{\pi} \sum_{k=1}^{N} \left| \gamma_k\left(k, \frac{k\pi}{2N}\right) \right| \cdot \left| v\left(t, \frac{k\pi}{2N}\right) - v(t, 0) \right| \tag{2.13}$$

$$E_3 \leqq \frac{2F}{\pi} \left| \sum_{k=1+N}^{\infty} \gamma_k(k, 0) \sin\left(\frac{k\pi x}{l}\right) v(t, 0) \right| \qquad 0 < x \leqq l \tag{2.14}$$

$$\gamma_k(k, y) = \frac{1}{k} y \cot y \qquad v(t, y) = \exp - \beta^2\pi^2k^2\left(\frac{\sin y}{y}\right)^2 \tag{2.15}$$

and $\beta^2 = \alpha t/l^2$ is a Fourier number.

Expressions (2.12) and (2.13) have been obtained as a consequence of the inequalities $|\sum a| \leqq \sum |a|$, the equation $|ab| = |a| \cdot |b|$, and following application of a mean value theorem. The decomposition of (2.11) into the forms shown is somewhat arbitrary. The present selection was largely dictated by the facility with which the resultant error components could be bounded. Conceptually, other arrangements of terms could lead to lower bounds on truncation error than those presented here.

3. Preliminary Results. In this section we shall prove some lemmas necessary for the comparison of μ and μ_h.

LEMMA 3.1. If $\gamma_k(k, y)$ is given by equation (2.15), then

$$\gamma_k(k, 0) - \gamma_k\left(k, \frac{k\pi}{2N}\right) \leqq kN^{-2}. \tag{3.1}$$

PROOF: Applying Eq. (2.15) we write

$$\gamma_k(k, 0) - \gamma_k(k, y) = \frac{1}{k}(1 - y \cot y). \qquad 0 < y \leq \pi/2, y = \frac{k\pi}{2N} \tag{3.2}$$

To equation (3.2) we apply the inequalities

$$0 \leqq \tfrac{1}{4}\pi^2(1 - y \cot y)y^{-2} \leqq 1 \tag{3.3}$$

to obtain

$$\gamma_k(k, 0) - \gamma_k\left(k, \frac{k\pi}{2N}\right) \leqq kN^{-2}$$

and the lemma is proved.

LEMMA 3.2. If $v(t, y)$ is defined as in Eq. (2.15) and $\beta > 0$, then

$$\left| v\left(t, \frac{k\pi}{2N}\right) - v(t, 0) \right| \leqq 8\beta^2k^3N^{-1} \exp - 4\beta^2k^2. \tag{3.4}$$

PROOF: In order to estimate the difference $v(t, k\pi/2N) - v(t, 0)$ with the aid of the theorem of the mean we have to estimate the right hand member of

the equation

$$(3.5)\quad \frac{\partial v}{\partial y}(t, y) = -2\beta^2\pi^2k^3y^{-1}\left(\frac{\sin y}{y}\right)\left[\cos y - \left(\frac{\sin y}{y}\right)\right]\exp - \beta^2\pi^2k^2\left(\frac{\sin y}{y}\right)^2.$$

$$0 \leqq y = \frac{k\pi}{2N} \leqq \frac{\pi}{2}$$

We observe that $\mu^{-1}\sin\mu$ is a non-increasing function of μ for $\mu > 0$ and write Jordan's inequality

$$(3.6)\qquad \frac{2}{\pi} \leqq \left(\frac{\sin y}{y}\right) \leqq 1 \qquad 0 \leqq y \leqq \pi/2.$$

In addition we write as a consequence of (3.6)

$$(3.7)\qquad \left(\frac{\sin y}{y}\right) \geqq \cos y \qquad 0 \leqq y \leqq \pi/2.$$

The second inequality follows from the relation $\tan\mu \geqq \sin\mu$ and (3.6). Therefore

$$(3.8)\quad 0 \leqq \frac{\partial v}{\partial y}(t, y) = 2\beta^2\pi^2k^2y^{-1}\left(\frac{\sin y}{y}\right)\left[\frac{\sin y}{y} - \cos y\right]\exp - \beta^2\pi^2k^2\left(\frac{\sin y}{y}\right)^2.$$

The right side of (3.8) assumes, in the interval for y, its maximum value at the upper limit, consequently

$$(3.9)\qquad 0 \leqq \frac{\partial v}{\partial y}(t, y) \leqq \frac{16}{\pi}\beta^2k^2\exp - 4\beta^2k^2.$$

Applying (3.9) to (3.8) and using a mean value theorem the lemma follows.

LEMMA 3.3. If $\gamma_k(k, y)$ is given by Eq. (2.15) and $1 \leqq k \leqq N$ then

$$(3.10)\qquad \left|\gamma_k\left(k, \frac{k\pi}{2N}\right)\right| \leqq \frac{2}{k}\left(1 - \frac{k^2}{N^2}\right).$$

PROOF: We estimate the coefficient $\gamma_k(k, k\pi/2N)$ by writing the inequality

$$(3.11)\qquad \tfrac{1}{6}\pi^3(1 - y\cot y)y^{-3} \geqq 1 \qquad 0 \leqq y \leqq \pi/2$$

where the equality holds at the upper limit of the interval. The inequality follows by noting that the left side takes on its minimum value at $y = \pi/2$.

Furthermore, from equations (2.15) there obtains

$$(3.12)\qquad |\gamma_k(k, 0)| = 1/k.$$

Application of (3.11) and (3.12) to (3.2) leads to the inequality

$$(3.13)\qquad \left|\gamma_k\left(k, \frac{k\pi}{2N}\right)\right| \leqq \frac{1}{k}\left(1 - \frac{k^3}{N^3}\right)$$

from which there follows the lemma by virtue of application of

$$(3.14)\qquad 1 - \frac{k^3}{N^3} \leqq 2\left(1 - \frac{k^2}{N^2}\right). \qquad 1 \leqq k \leqq N.$$

Finally we present the following three lemmas related to bounds and a mean value property of sums.

LEMMA 3.4. If a is non-negative, then the probability exponential sum can be bounded above by

$$\sum_{k=1}^{M} k \exp - ak^2 \leqq \frac{1}{2a}\left[\sqrt{\pi a}\ \operatorname{erf} M\sqrt{a} + (1 - \exp - aM^2)\right]. \tag{3.15}$$

PROOF: We consider the identity

$$\sum_{k=1}^{M} k \exp - ak^2 = \int_1^{M+1} v \exp - av^2\, dv + \sum_{k=1}^{M} \int_k^{k+1} [k \exp - ak^2 - v \exp - av^2]\, dv \tag{3.16}$$

$$k < v < k + 1; v - 1 < k < v.$$

Now note that the second integrand on the right of Eq. (3.16) can be bounded above by

$$k \exp - ak^2 - v \exp - av^2 \leqq v \exp - a(v-1)^2 - v \exp - av^2 \tag{3.17}$$

consequently Eq. (3.16) can be written as

$$\sum_{k=1}^{M} k \exp - ak^2 \leqq \int_1^{M+1} v \exp - av^2\, dv + \int_1^{M+1} v[\exp - a(v-1)^2 - \exp - av^2]\, dv \tag{3.18}$$

and the lemma is proved.

LEMMA 3.5. If $G(k)$ is stepwise continuous and non-negative in $0 \leqq x \leqq N$ and if the sign of $G'(x)$, where $G(k)$ is the bounded interpolating function on $G(x)$, changes from positive to negative P times in $0 \leqq x \leqq N$, then the sum of functions $G(k)$ is expressed in terms of its Riemann integral as

$$\sum_{k=1}^{N} G(k) \leqq \int_0^{N+1} G(x)\, dx + \sum_{p=1}^{P} G(X_p) \qquad 0 \leqq X_p \leqq N \tag{3.19}$$

where X_p is the value of x at the location of the p'^{th} sign change in $G'(x)$.

PROOF: The lemma can be demonstrated by consideration of the area bounded by $G(x)$.

LEMMA 3.6. If $G(x)$ and $H(x)$ are continuous functions for $0 \leqq x \leqq M$, and $H(x)$ is non-negative, then

$$|\textstyle\sum_{k=1}^{M} G(k)H(k)| \leqq H\,|\sum_{k=1}^{M} G(k)| \tag{3.20}$$

where H is the max of H in $1 \leqq k \leqq M$.

PROOF: Lemma 3.6 is a special case of Holder's inequality, considered in Hardy's *Inequalities*, p. 25.

4. The Truncation Error. Our preceding results can now be collected into a truncation error theorem.

THEOREM 4.1. The solution of the semi-discrete problem (2.1) in the open rectangle $0 < x \leqq l$, $t > 0$, subject to boundary and initial conditions (2.2) and (2.3) differs in absolute value from the differential solution to the same problem (2.10) by the truncation error

$$\begin{aligned}\frac{\pi}{F}\,|\,\mu - \mu_h\,| \leqq{}& \frac{1}{(\pi A)^2}\left[\pi\sqrt{\pi}\,\frac{A}{N}\operatorname{erf}\pi A + (1 - \exp - \pi^2A^2)\right] \\ &+ \frac{\sqrt{\pi}}{A}\left(1 - \frac{3}{8A^2}\right)\operatorname{erf} 2B \\ &+ \left[4\,\frac{B}{A}\left(\frac{M^2}{N^2} - 1\right) + \frac{3}{2}\,\frac{B}{A^3}\right]\exp - 4B^2 + \frac{8}{N}\,[(1 + 2A^2) - \sqrt{1 + 4A^4}] \\ &\cdot\left[1 - \frac{1}{4A^2}\,\{(1 + 2A^2) - \sqrt{1 + 4A^4}\}\right] \\ &\cdot \exp - [(1 + 2A^2) - \sqrt{1 + 4A^4}] + \pi \exp - \pi^2A^2\left(1 + \frac{1}{N}\right)^2\end{aligned} \tag{4.1}$$

where $A = \beta N$, $B = \beta M$, $M = N + 1$.

PROOF: Consider first the definition (2.12) of E_1. By lemma 3.1 and definition (2.15) this reduces to

$$E_1 \leqq \frac{2F}{\pi}\sum_{k=1}^{N}\frac{k}{N^2}\exp - \beta^2\pi^2k^2. \tag{4.2}$$

Upon applying lemma 3.4 we have

$$\frac{\pi E_1}{F} \leqq \frac{1}{(\pi A)^2}\left[\frac{A\pi}{N}\sqrt{\pi}\operatorname{erf}\pi A + (1 - \exp - \pi^2A^2)\right]. \tag{4.3}$$

Now refer to the definition of E_2, Eq. (2.13). Applying lemmas 3.2, 3.3 to it we have

$$\frac{\pi}{F}E_2 \leqq \frac{32A^2}{N^3}\sum_{k=1}^{N}k^2\left(1 - \frac{k^2}{N^2}\right)\exp - 4\beta^2k^2. \tag{4.4}$$

Finally we apply lemma 3.5 to bound Eq. (4.4) above by

$$\begin{aligned}\frac{\pi}{F}E_2 \leqq{}& \frac{\sqrt{\pi}}{A}\left(1 - \frac{3}{8A^2}\right)\operatorname{erf} 2B + \frac{3}{2}\,\frac{B}{A^3}\exp - 4B^2 \\ &+ 4\,\frac{B}{A}\left(\frac{M^2}{N^2} - 1\right)\exp - 4B^2 + \frac{8}{N}\,[(1 + 2A^2) - \sqrt{1 + 4A^4}] \\ &\cdot\left[1 - \{(1 + 2A^2) - \sqrt{1 + 4A^4}\}\,\frac{1}{4A^2}\right]\exp - [(1 + 2A^2) - \sqrt{1 + 4A^4}]\end{aligned} \tag{4.5}$$

since of the locations of sign change of $G'(x)$

$$X_1 = \frac{N}{2A}\sqrt{(1+2A^2)-\sqrt{1+A^4}} < M;$$

$$X_2 = \frac{N}{2A}\sqrt{(1+2A^2)+\sqrt{1+A^4}} > M. \tag{4.6}$$

To bound E_2 above we have by virtue of (2.14) and (2.15)

$$\frac{\pi E_2}{F} \leqq 2\left|\sum_{k=1+N}^{\infty}\frac{1}{k}\sin\left(\frac{k\pi x}{l}\right)\exp -\beta^2\pi^2k^2\right| \qquad 0 < x \leqq l. \tag{4.7}$$

To this result we apply lemma 3.6. There results

$$\frac{\pi E_2}{F} \leqq 2\left|\sum_{k=1+N}^{\infty}\frac{1}{k}\sin\left(\frac{k\pi x}{l}\right)\right|\exp -\pi^2A^2\left(1+\frac{1}{N}\right)^2. \tag{4.8}$$

The right side of (4.8) can be further reduced to

$$\frac{\pi E_2}{F} \leqq 2\left|\frac{\pi}{2}-\int_0^{\pi x/l}\frac{\sin\,(N+\frac{1}{2})t}{2\sin\frac{1}{2}t}\,dt\right|\exp -\pi^2A^2\left(1+\frac{1}{N}\right)^2$$

hence

$$\frac{E_2}{F} \leqq \exp -\pi^2A^2\left(1+\frac{1}{N}\right)^2 \tag{4.9}$$

as a consequence of the result found in Courant (18).

Expression (4.9) is a greater upper bound than would be used in the evaluation of local error. In fact, examination of (4.8) reveals that E_2 vanishes at the right boundary of the interval $(0, l)$.

Adding the three inequalities (4.3), (4.5), (4.9) we complete the proof of the theorem.

The convergence of the result (4.1) for large N may be demonstrated by putting $N = l/h$ and passing to the limit as $h \to 0$. The convergence of lowest order is thereby seen to be $O(1/N)$. This corresponds to the asymptotic value of the truncation error for large A, namely

$$\frac{1}{F}\,|\,\mu - \mu_h\,| \leqq \frac{8}{\pi eN}.$$

It can be noted that lemma 3.5 can be applied to (4.2) instead of lemma 3.4. The resultant bound is slightly lower than Eq. (4.3) at the expense of introducing two expressions for E_1 in the intervals for A considered. In a similar fashion the approach offered by Eq. (3.16) leading to lemma 3.4, when applied to Eq. (4.4) yields a greater bound for E_2 than that available through lemma 3.5.

Passing to the limit as $\beta \to 0$ we find for the truncation error

$$|\,\mu - \mu_h\,| \leqq \frac{F}{\pi}\left(1+\pi+\frac{2}{N}\right). \tag{4.10}$$

This is slightly larger than the maximum value obtained directly from (4.2), (4.4) and (4.7),

$$|\mu - \mu_h| \leqq \frac{F}{\pi}\left(1 + \pi + \frac{1}{N}\right) \tag{4.11}$$

by applying the sum

$$\sum_{k=1}^{N} \frac{k}{N^2} = \frac{1}{2}\left(1 + \frac{1}{N}\right). \tag{4.12}$$

Numerical results pertaining to Theorem 4.1 are presented in figure 1. The relative minimum, in the results, for each value of N, is also found in the results of Klein, et al. Since the maximum value of $|\mu - \mu_h|$ is unity, the data have been restricted accordingly. The fact that, for large N, μ_h does not converge to μ is a consequence of our method of estimating global error. A comparison between (2.4) and (2.10) shows that $\lim_{N\to\infty} \mu_h = \mu$.

5. Heat Truncation Error. The preceding results can be utilized to obtain an error bound on heat flux $q = k\mu_x$, instead of temperature, μ, in a problem with modified boundary conditions. To this end we refer again to the heat equation (2.7), and state *a*.

THEOREM 5.1. For the problem (2.1)–(2.3) the truncation error in $q = k\mu_x$ namely $|q - q_h|$, associated with the solution of $\nabla^2 \times q = (1/\alpha)(\partial q/\partial t)$ and its semi-discrete analog, subject to the boundary and initial conditions

$$\mu(x, 0) = q(x, 0) = 0, \qquad q(0, t) = F, \qquad q(l, t) = 0 \tag{5.1}$$

is given by relation (4.1) with μ replaced by q.

PROOF: The proof follows by differentiating (2.7) with respect to x, introducing $q = k\mu_x$, and replacing μ_x by q.

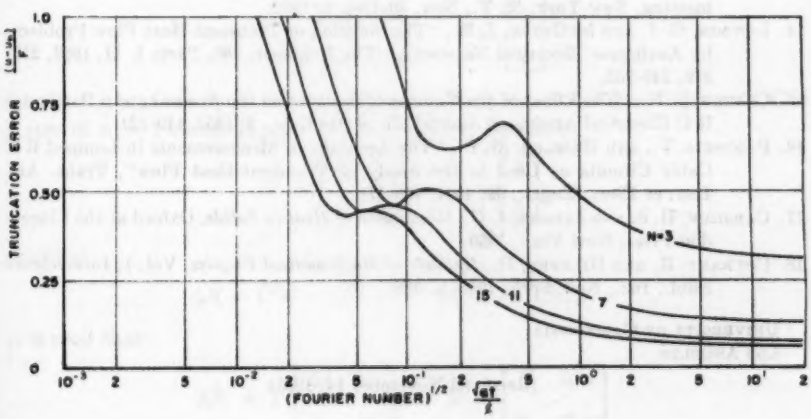


FIG. 1

REFERENCES

1. JOHN, F., "On Integration of Parabolic Equations by Difference Methods", Part I, "Linear and Quasi-linear Equations for the Infinite Interval", Commun. on Pure and Appl. Math., *5*, 1952, 155–211.
2. O'BRIEN, G. G., HYMAN M. A., AND KAPLAN, S., "A Study of the Numerical Solution of Partial Differential Equations", Jl. of Math. and Phys., *29*, 1951, 223–251.
3. JUNCOSA, M. L. AND YOUNG, D. M., "On the Order of Convergence of Solutions of a Difference Equation to a Solution of the Diffusion Equation", Jl. of the Soc. for Indust. and Appl. Math., *1*, 1953, 111–135.
4. LEUTERT, W., "On the Convergence of Unstable Approximate Solutions of the Heat Equation to the Exact Solution", Jl. of Math. and Phys., *29*, 1951, 245–251.
5. HILDEBRAND, F. B., "On the Convergence of Numerical Solutions of the Heat Flow Equation", Jl. of Math. and Phys., *31*, 1952, 35–41.
6. BLANCH, G., "On the Numerical Solution of Parabolic Partial Differential Equations", Jl. of Res., Nat. Bur. of Stand., *50*, 1953, 343–356.
7. CRANDALL, S. T., "Implicit vs Explicit Recurrence Formulas for the Linear Diffusion Equation". Jl. of the Assoc. for Computing Mach., *2*, 1955, 42–49.
8. CRANK, J. AND NICOLSON, P., "A Practical Method for Numerical Evaluation of Partial Differential Equations of the Heat Conduction Type," Proc. Cambr. Philos. Soc., *43*, 1947, 50–67.
9. HILDEBRAND, F. B., *Methods of Applied Mathematics*, Prentice-Hall, Inc., 1956, P. 379, Prob. 98.
10. HOWE, R. M., AND HANEMAN, V S., "The Solution of Partial Differential Equations by Difference Methods Using the Electronic Differential Analyzer", Proc. of the Inst. of Radio Engrs., Computer Issue, 1953, 1497–1508.
11. HARTREE, D. R. AND WOMERSLEY, J. R., "A Method for the Numerical or Mechanical Solution of Certain Types of Partial Differential Equations", Proc. Roy. Soc. London (A), *161*, 1937, 353–366.
12. PASCHKIS, V. AND BAKER, H. D., "A Method for Determining Unsteady State Heat Transfer by Means of an Electrical Analogy", Trans. Am. Soc. Mech. Engrs., *64*, 1942, 105.
13. KLEIN, E. O. P., TOULOUKIAN, Y. S. AND EATON, J. R., "Limits of Accuracy of Electrical Analog Circuits Used in the Solution of Transient Heat Conduction Problems", Am. Soc. of Mech. Engrs., Paper No. 52-A-65, Presented at annual meeting, New York, N. Y., Nov. 30–Dec. 5, 1952.
14. LAWSON, G. I. AND MCGUIRE, J. H., "The Solution of Transient Heat Flow Problems by Analogous Electrical Networks," The Engineer, *196*, Parts I, II, 1953, 217–219, 249–252.
15. CLARKE, L. N., "The Effect of the Number of Sections on the Accuracy of a Particular R C Electrical Analogue, Austral. Jl. of Appl. Sc., *3*, 1952, 119–124.
16. PASCHKIS, V., AND HEISLER, M. P., "The Accuracy of Measurements in Lumped R-C Cable Circuits as Used in the Study of Transient-Heat Flow", Trans. Am. Inst. of Elect. Engrs., *63*, 1944, 165–171.
17. CARSLAW, H. S. AND JAEGER, J. C., *Conduction of Heat in Solids*, Oxford at the Clarendon Press, New York, 1950.
18. COURANT, R. AND HILBERT, D., *Methods of Mathematical Physics*, Vol. 1, Interscience Publ., Inc., New York, 1953, p. 106.

UNIVERSITY OF CALIFORNIA,
LOS ANGELES

(Received November 14, 1955)

NOTE ON THE SENSITIVITY OF LEAST SQUARES SOLUTIONS

By Mark Lotkin

In solving problems related to the adjustment of observational data by least squares methods, there frequently arises the question of the effect of observational errors upon the solution. The answer can be supplied readily for the general case:

Let

$$Cx = f \tag{1}$$

represent the set of conditional equations, with C denoting a matrix of r rows and n ($r \geqq n$) columns. The associated set of normal equations is then

$$Nx_e = g, \tag{2}$$

with $N = C^TC$, $g = C^Tf$, and the superscript T denoting transposition.

Let us assume now that C and b are subjected to certain changes, leading to a new set of observational equations

$$\bar{C}\bar{x} = \bar{f}. \tag{3}$$

Of interest, then, is the difference $\Delta x = \bar{x}_e - x_e$. Clearly,

$$\Delta x = \bar{N}^{-1}(\Delta g - \Delta N x_e), \tag{4}$$

with

$$\Delta g = \bar{g} - g, \Delta N = \bar{N} - N.$$

We shall restrict ourselves here to the consideration of a particular case, namely the replacement of only one conditional equation by another, linearly independent one. Let us first partition C, f, respectively, as

$$C = \begin{bmatrix} A & b \\ c & \alpha \end{bmatrix}, \qquad f = \begin{bmatrix} h \\ \phi \end{bmatrix},$$

where A is an $(r-1) \times (n-1)$ matrix, b, h are columns of $r-1$ elements, C is a row of $n-1$ elements, and α, ϕ are scalars. Similarly,

$$\bar{C} = \begin{bmatrix} A & b \\ \bar{c} & \bar{\alpha} \end{bmatrix}, \qquad \bar{f} = \begin{bmatrix} h \\ \bar{\phi} \end{bmatrix}.$$

Writing

$$\Delta N = \bar{C}^T\bar{C} - C^TC = \bar{C}^T\,\Delta C + (\Delta^T C)C,$$

it is seen that

$$\Delta N = T\begin{bmatrix} \Delta c & \Delta\alpha \\ c & \alpha \end{bmatrix}, \qquad T = \begin{bmatrix} \bar{c}^T & \Delta^T c \\ \bar{\alpha} & \Delta\alpha \end{bmatrix}.$$

Further,

$$\Delta g = \bar{C}^T \Delta f + \Delta^T C f = T \begin{bmatrix} \Delta\phi \\ \phi \end{bmatrix}.$$

Consequently,

$$\Delta x = \bar{N}^{-1} T \left(\begin{bmatrix} \Delta\phi \\ \phi \end{bmatrix} - \begin{bmatrix} \Delta c & \Delta\alpha \\ c & \alpha \end{bmatrix} x_c \right). \tag{5}$$

The main problem lies in the calculation of $\bar{N}^{-1}$. In certain cases this computation may be reduced greatly. This is the case, for example, if $\bar{c} = c$. Then, clearly, N and $\bar{N}$ may be partitioned as:

$$N = \begin{bmatrix} P & q \\ q^T & \rho \end{bmatrix}, \qquad \bar{N} = \begin{bmatrix} P & \bar{q} \\ \bar{q}^T & \bar{\rho} \end{bmatrix}.$$

Thus, if

$$\bar{N}^{-1} = \begin{bmatrix} R & s \\ s^T & \tau \end{bmatrix}$$

then the parts τ, s, R of $\bar{N}^{-1}$ may be obtained by the well known algorithm

$$Y = P^{-1}\bar{q}, \quad \theta = \bar{\rho} - \bar{q}^T Y, \quad \tau = 1/\theta, s = -Y\tau, \quad R = P^{-1} + Y(Y^T\tau). \tag{6}$$

Knowing P^{-1} this algorithm requires $(n-1)(3n+2)/2$ multiplications for the construction of $\bar{N}^{-1}$, while it takes considerably more multiplications if one were to perform the total inversion.

The further assumption that also $\bar{\alpha} = \alpha$ leads to the simple relationship

$$\Delta x = N^{-1} \begin{bmatrix} c^T \\ \alpha \end{bmatrix} \Delta\phi. \tag{7}$$

It permits the determination of the effect upon the solution x of an error $\Delta\phi$ in a conditional equation.

It is of interest to calculate also the change in the residual vector $r = Cx_c - f$. Clearly,

$$\bar{r}^T\bar{r} = \bar{x}_c^T N \bar{x}_c - 2\bar{x}_c^T \bar{g} + \bar{f}^T\bar{f}.$$

Putting $\bar{x}_c = x_c + \Delta x$, with Δx given by (7), and using the abbreviation $u^T = (c, \alpha)$, there results for the change in length of r the expression

$$\Delta r^2 \equiv \bar{r}^T\bar{r} - r^T r = [u^T N^{-1}(u\Delta\phi - 2\bar{g}) + (\bar{\phi} + \phi)]\Delta\phi \tag{8}$$

Let us illustrate above conclusions by a simple example. Suppose, then, measurements f taken at abscissas "a" result in the following table:

a	f
19	.80
20	.92
21	1.00

It is desired to fit a straight line

$$x_1 a + x_2 = f$$

to the given points. Here

$$C = \begin{bmatrix} 19 & 1 \\ 20 & 1 \\ 21 & 1 \end{bmatrix} \qquad f = \begin{bmatrix} .80 \\ .92 \\ 1.00 \end{bmatrix}.$$

Further,

$$N = \begin{bmatrix} 1202 & 60 \\ 60 & 3 \end{bmatrix}, \qquad N^{-1} = \frac{1}{6}\begin{bmatrix} 3 & -60 \\ -60 & 1202 \end{bmatrix},$$

and

$$x_e = \frac{1}{600}\begin{bmatrix} 60 \\ -656 \end{bmatrix}.$$

Thus

$$r = Cx_e - f = \frac{1}{150}\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}, \quad \text{and} \quad r^T r = (16/6)\ 10^{-4}.$$

Suppose, now, that the measurement of $\phi = f_3 = 1.00$ was too low by 1%, so that $\bar{f}_3 = 1.01$, $\Delta\phi = 0.01$. By (7), then,

$$\Delta x = \frac{1}{600}\begin{bmatrix} 3 \\ -58 \end{bmatrix} \qquad \bar{x}_e = \frac{1}{600}\begin{bmatrix} 63 \\ -714 \end{bmatrix}.$$

The change of 1% in f_3 thus causes a change of 5% and 9% in the components of x_e.

Making use of (8) to compute the change in $\bar{r}^T r$ we obtain

$$\Delta r^2 = -(7/6)\ 10^{-4}$$

$$\bar{r}^T \bar{r} = (3/2)\ 10^{-4}.$$

Thus the change of 1% in one of the observations reduces the square of the length of the residual vector by about 44%.

AVCO RESEARCH DIVISION

(Received September 26, 1955)

A VECTOR METHOD FOR SOLVING LINEAR EQUATIONS AND INVERTING MATRICES

By J. P. Roth* and D. S. Scott†

Introduction. In a recent article E. W. Purcell (1) presented a vector method for finding solutions of simultaneous linear equations; the method he described does not always work, however, since division by zero may occur. In this paper we remedy this defect, add a simplification and extend the method to matrix inversion. Also we present a simple scheme for solving equations with paper and pencil or desk computer, having the advantage over Cramer's method, for instance, that one's work may be checked at each stage of the computation. Some advantages and disadvantages of the method of matrix inversion are discussed: An important advantage from the standpoint of digital machine calculation is that it permits rescaling at each stage which makes the method less sensitive to round off errors than several other methods; a disadvantage is its relative slowness, requiring roughly $5n^3$ operations to invert an $n \times n$ matrix. In another paper we shall show how this can be obtained.

1. Vector interpretation of solving systems of equations. Consider a system of equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n + b_1 &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_nx_1 + \cdots + a_{nn}x_n + b_n &= 0 \end{aligned} \tag{1}$$

This system defines the set of n row vectors in $(n + 1)$ dimensional Euclidean space

$$E_i = (a_{i1}, a_{i2}, \cdots, a_{in}, b_i) \qquad i = 1, \cdots, n \tag{2}$$

The system (1) may then be interpreted as a statement that the vector $(x_1, \cdots, x_n, 1)$ is orthogonal to each of the E_i. The problem of solving (1) may then be stated as that of finding such a vector which is orthogonal to each of the row vectors E_i; more generally the problem is to find a basis for vectors of this type which are orthogonal to all the E_i.

We shall present a method for constructing such a basis, the method depending upon the following lemma.

2. Lemma. Let X be a subspace of a vector space V, $u_1, \cdots, u_m$ a basis for X and E a vector in V which is not orthogonal to u_1, $E \cdot u_1 \neq 0$. Then the vectors

$$\begin{aligned} v_1 &= u_1(u_2 \cdot E) - u_2(u_1 \cdot E) \\ v_2 &= u_1(u_3 \cdot E) - u_3(u_1 \cdot E) \\ &\vdots \\ v_{m-1} &= u_1(u_m \cdot E) - u_m(u_m \cdot E) \end{aligned}$$

* Now at The Institute for Advanced Study, Princeton, New Jersey
† Now at Princeton University, Princeton, New Jersey

are (1) orthogonal to E, (2) linearly independent, and (3) form a basis for the subspace Y of X which is orthogonal to E.

Proof: By direction computation $E \cdot v_j = (E \cdot u_1)(u_{j+1} \cdot E) - (E \cdot u_{j+1})(u_1 \cdot E) = 0, j = 1, \cdots, m-1$. Let

$$v_m = u_1 .$$

Let us write out the linear transformation defined in the above equations as a matrix equation

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{m-1} \\ v_m \end{pmatrix} = \begin{pmatrix} u_2 \cdot E, & -u_1 \cdot E, & 0, & \cdots & 0, & 0 \\ u_3 \cdot E, & 0, & -u_1 \cdot E, & 0, \cdots & 0, & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ u_m \cdot E, & 0, & 0, & \cdots & 0, & -u_1 \cdot E \\ 1, & 0, & 0, & \cdots & 0, & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_m \end{pmatrix}$$

It can be seen immediately that the transformation matrix is nonsingular, for expanding the determinant by the last row we obtain $\pm(u_1 \cdot E)^m$, and $u_1 \cdot E$ is not zero by hypothesis. Hence the inverse transformation exists so that $u_1, \cdots, u_m$ can be expressed as linear combination of the vectors $v_1, \cdots, v_m$. Thus $v_1, \cdots, v_m$ constitute a basis for X. It easily follows that $v_1, \cdots, v_{m-1}$ form a basis for the subspace Y of X orthogonal to E, and Y is $(m-1)$ dimensional, q.e.d.

3. Construction of the solution by the vector method. Referring to equations (1) and (2), we start with the following basis of unit vectors for Euclidean $(n+1)$ space: $e_1 = (1, 0, 0, \cdots, 0)$, $e_2 = (0, 1, 0, \cdots, 0), \cdots, e_{n+1} = (0, 0, \cdots, 1)$. We form the inner products $e_1E_1, e_2E_1, \cdots, e_{n+1}E_1$ and assume without loss of generality that $e_1E_1 \neq 0$. Then by the above lemma the following set

$$f_i = (e_{i+1}E_1)e_i - (e_iE_1)e_{i+1}, \qquad i = 1, \cdots, n$$

of n vectors are orthogonal to E_1 and in fact form a basis for the n-dimensional subspace orthogonal to E_1. Now we form the inner products $f_1E_2, f_2E_2, \cdots, f_nE_2$. If all of these are zero then E_1 and E_2 are linearly dependent, the f's automatically satisfy E_2 and we may pass on directly to E_3. If one of these products is not zero, say f_1E_2, then the $n-1$ vectors

$$g_j = (f_{j+1}E_2)f_1 - (f_1E_1)f_{j+1} \qquad j = 1, \cdots, n-1$$

are orthogonal to E_2 and, being linear combinations of the f_i, are also orthogonal to E_1 and form a basis for the orthogonal complement of the subspace spanned by E_1 and E_2.

Proceeding inductively, we suppose that

$$g_1, \cdots, g_{n+1-k}$$

is a basis for the subspace in E^{n+1} Euclidean $(n+1)$ space orthogonal to the subspace spanned by

$$E_1, \cdots, E_k \qquad 1 \leqq k < n.$$

We proceed to E_{k+1} ; if $g_i E_{k+1} = 0$ for all g_i then we already have a basis orthogonal to the subspace generated by $E_1, \cdots, E_{k+1}$. If, however, say $g_1 \cdot E_{k+1} \neq 0$ then

$$h_r = (g_{r+1} \cdot E_{k+1}) g_1 - (g_1 \cdot E_{k+1}) g_{r+1}, \cdots, r = 1, \cdots, n - k$$

forms a basis orthogonal to the subspace spanned by $E_1, \cdots, E_{k+1}$. Thus, proceeding inductively we obtain a basis for the subspace orthogonal to all the E_i, or we exhaust the n-dimensional subspace before we exhaust the E_i, in which case there is no solution. Thus either we obtain a basis for all solutions of the equations (1) or else no solution exists. We will now consider a numerical example.

4. A scheme for systematizing the solution of linear equations. We shall now outline a "*rote*" *method* for solving equations, which can easily be carried out with paper and pencil with or without a desk computer for not too large systems of equations. We shall work out an example for 5 equations in 5 unknowns. Consider then the system

$$\begin{aligned}
x_1 + 2x_2 \quad\quad + 2x_4 + \ x_5 + 0 &= 0 \\
3x_1 + 2x_2 + \ x_3 + 2x_4 + 3x_5 + 0 &= 0 \\
x_1 + 3x_2 + 5x_3 + \ x_4 + \ x_5 + 2 &= 0 \\
x_1 \quad\quad - \ x_4 + \ x_5 + 1 &= 0 \\
2x_1 + 2x_2 + \ x_3 + \ x_4 + 3x_5 - 1 &= 0
\end{aligned} \tag{3}$$

Table I will be almost self-explanatory:

1. First write down the row vectors $E_1, \cdots, E_5$ defined by (3), underneath these, the unit vectors $e_1, \cdots, e_5$;
2. Then in a single column just to the right of each e vector separated from it by bold lines we record the products of E_1 with each e vector.
3. We encircle any one of these numbers which is not zero; in this example we choose $e_1 \cdot E_1$, and *change its sign.*
4. We then compute a new vector f_1 by multiplying e_1 by the number $(e_2 \cdot E_1)$ opposite e_2 and adding this vector to the vector e_2 multiplied by the (new) number opposite e_1,

$$\begin{array}{lc} (1,0,0,0,0) & 1^- \\ & \times \\ (0,1,0,0,0) & 2 \end{array}$$

We use this "criss-cross" method to compute the remaining f's which are recorded directly below the e's (cf. adjacent formulas for the f's).

5. We form the products of E_2 with the f's and record these in the column just adjacent to the f's.
6. We encircle one of these which is nonzero (in this example we choose $f_2 \cdot E_2$) and *change its sign.*

TABLE I

Scheme for Solving a System of Five Equations in Five Unknowns

E_1	1	2	0	2	1	0		
E_2	3	2	1	2	3	0		
E_3	1	3	5	1	1	2		
E_4	1	0	0	−1	1	1		
E_5	2	2	1	1	3	−1		
e_1	1	0	0	0	0	0	①⁻	$(=e_1\cdot E_1)$
e_2	0	1	0	0	0	0	2	$(=e_2\cdot E_1)$
e_3	0	0	1	0	0	0	0	$(=e_3\cdot E_1)$
e_4	0	0	0	1	0	0	2	$(=e_4\cdot E_1)$
e_5	0	0	0	0	1	0	1	$(=e_5\cdot E_1)$
e_6	0	0	0	0	0	1	0	$(=e_6\cdot E_1)$
f_1	2	−1	0	0	0	0	4	$(=f_1\cdot E_2)$
f_2	0	0	−1	0	0	0	−①⁺	$(=f_2\cdot E_2)$
f_3	2	0	0	−1	0	0	4	$(=f_3\cdot E_2)$
f_4	1	0	0	0	−1	0	0	$(=f_4\cdot E_2)$
f_5	0	0	0	0	0	−1	0	$(=f_5\cdot E_2)$
g_1	2	−1	−4	0	0	0	−21	$(=g_1\cdot E_3)$
g_2	2	0	−4	−1	0	0	−19	$(=g_2\cdot E_3)$
g_3	1	0	0	0	−1	0	0	$(=g_3\cdot E_3)$
g_4	0	0	0	0	0	−1	−②⁺	$(=g_4\cdot E_3)$
h_1	4	−2	−8	0	0	21	+㉙⁻	$(=h_1\cdot E_4)$
h_2	4	0	−8	−2	0	19	25	$(=h_2\cdot E_4)$
h_3	2	0	0	0	−2	0	0	$(=h_3\cdot E_4)$
i_1	0	−2	0	2	0	2	−④⁺	$(=i_1\cdot E_5)$
i_2	−2	0	0	0	2	0	2	$(=i_2\cdot E_5)$
j	−4	−2	0	2	4	2		

$f_1 = e_1(e_2\cdot E_1) - e_2(e_1\cdot E_1)$

$f_2 = e_1(e_3\cdot E_1) - e_3(e_1\cdot E_1)$

$f_3 = e_1(e_4\cdot E_1) - e_4(e_1\cdot E_1)$

$f_4 = e_1(e_5\cdot E_1) - e_5(e_1\cdot E_1)$

$f_5 = e_1(e_6\cdot E_1) - e_6(e_1\cdot E_1)$

$g_1 = f_2(f_1\cdot E_2) - f_1(f_2\cdot E_2)$

$g_2 = f_2(f_3\cdot E_2) - f_3(f_2\cdot E_2)$

$g_3 = f_2(f_4\cdot E_2) - f_4(f_2\cdot E_2)$

$g_4 = f_2(f_5\cdot E_2) - f_5(f_2\cdot E_2)$

$h_1 = g_4(g_1\cdot E_3) - g_1(g_4\cdot E_3)$

$h_2 = g_4(g_2\cdot E_3) - g_2(g_4\cdot E_3)$

$h_3 = g_4(g_3\cdot E_3) - g_3(g_4\cdot E_3)$

$i_1 = h_1(h_2\cdot E_4) - h_2(h_1\cdot E_4)\,(1/25)$

$i_2 = h_1(h_3\cdot E_4) - h_3(h_1\cdot E_4)\,(1/25)$

$j_1 = i_2(i_1\cdot E_5) - i_1(i_2\cdot E_5)\quad (1/2)$

Answer: $x_1 = -4/2 = -2$

$x_2 = -2/2 = -1$

$x_3 = 0/2 = 0$

$x_4 = 2/2 = 1$

$x_5 = 4/2 = 2$

(We divided i_1, i_2 by 25 and j_1 by 2 merely to reduce the size of the numbers.)

7. We then use the criss-cross method to form the g's (cf. adjacent formulas).
8. And so on until we arrive at j.
9. Finally, we divide j by its last component, which will always be nonzero if there is a solution to the original system of equations, to give the solution

5. Solving matrix equations. We shall indicate by means of an example how the method is extended to inverting matrices. Consider the matrix equation

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x & u & p \\ y & v & q \\ z & w & v \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4}$$

Considering just the first column of each side of the equation, we have the equations

$$a_{11}x + a_{12}y + a_{13}z - 1 = 0 \tag{5}$$

$$a_{21}x + a_{22}y + a_{23}z - 0 = 0 \tag{6}$$

$$a_{31}x + a_{32}y + a_{33}z - 0 = 0 \tag{7}$$

which, considering x, y, z as unknowns, we see to be a special case of equation (1). We could therefore solve this equation by the method outlined in article 1, but it may be done even more simply by taking advantage of the zeros appearing in the first column of the identity matrix of equation (4); we interpret equations (6) and (7) as stating that the 3-vector (x, y, z) is orthogonal to the row vectors $E_2 = (a_{21}, a_{22}, a_{23})$ and $E_3 = (a_{31}, a_{32}, a_{33})$. Then consider unit vectors $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$. We now form $e_1 \cdot E_3$, $e_2 \cdot E_3$, $e_3 \cdot E_3$. At least one of these products will be non-zero unless $E_3 = (0, 0, 0)$, which case we rule out. Suppose then that $e_1 \cdot E_3 \neq 0$. We form

$$f_1 = (e_2 \cdot E_3)e_1 - (e_1 \cdot E_3)e_2$$

$$f_2 = (e_3 \cdot E_3)e_1 - (e_1 \cdot E_3)e_3 .$$

Then $f_1 \cdot E_3 = f_2 \cdot E_3 = 0$. Next we form $f_1 \cdot E_2$ and $f_2 \cdot E_2$. Again not both of these will be zero and $g = (f_1E_2)f_2 - (f_2E_2)f_1$ is orthogonal to E_3 (and, of course, E_3). Again, g will not be orthogonal to $E_1 = (a_{11}, a_{12}, a_{13})$ when the matrix (a_{ij}) is nonsingular. Then $\bar{g} = (x, y, z) = 1(E_1 \cdot g)g$ will be a column vector such that $\bar{g} \cdot E_1 = 1$, $\bar{g} \cdot E_2 = 0 = \bar{g} \cdot E_3$. This procedure then gives us the first column of our inverse. A similar technique can be used to get the second and third columns. Our work is shortened by the fact, however, that in finding the second column for the inverse we find a vector orthogonal to E_3 and E_1, but in our computations for the first column we already found a basis f_1, f_2, equations (8), orthogonal to E_3 and thus a vector orthogonal to E_3 and E_1 can be written down at once, $h = f_1(E_1f_2) - f_2(E_1f_1)$. Then if h has the components (h_1, h_2, h_3), then the second column of our inverse will be $(u, v, w) = (h_1/E_2 \cdot h, h_2/E_2 \cdot h, h_3/E_3 \cdot h)$. One obtains the third column of the inverse in the same manner as the first.

The procedure for inverting higher order matrices is essentially the same; the shortcut discussed above, namely of using information from previously solved columns to reduce work at a later stage, becomes much more prominent however.

5. An evaluation of the method.

For detailed information on and comparisons of various other methods for

inverting matrices we refer the reader to references 2, 3 and 4. We list here what we consider to be the important advantages and disadvantages of the vector method as compared with other methods.

(1) The vector method requires no special properties of the matrix (such as having the elements on the main diagonal large in comparison with the off-diagonal elements; the Crout method, for instance, has this requirement). (2) Perhaps, from the standpoint of its adaptability to machine operation, its chief merit is that it allows for frequent rescaling throughout the computation. We have encountered examples, in radiation analysis, of matrices where both the Crout and Gauss-Seidel methods failed on a CPC, the first because of scaling difficulties and the second because of extreme slowness of convergence. The vector method, however, was successful. (3) No division occurs except at the end. (4) It affords a convenient method for solving equations by hand. (5) For solving equations it requires less storage than the Crout method.

An important disadvantage is that it is slow in comparison with many other methods. It requires roughly $5n^3$ multiplications to invert an $n \times n$ matrix, whereas other direct methods are of the order of n^3. In a subsequent paper we shall show how this can be obtained.

BIBLIOGRAPHY

1. Purcell, E. W., "The Vector Method of Solving Linear Equations," J. of Math. and Phys., *32* (1953), 180–183.
2. Fox, L., "Practical Solutions of Linear Equations and Inversion of Matrices," Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues, National Bureau of Standards, Applied Mathematics Series No. 39, 1954.
3. Trimble, G. R., Jr., "A Survey of Methods of Inverting Matrices and Solving Systems of Simultaneous Linear Algebraic Equations," IBM Report 101 001 282 (1953) This report gives a large bibliography.
4. Bodewig, E., "Bericht über die verschiedenen Methoden zur Lösung eines Systems linearer Gleichungen mit reelen Koeffizienten," Nederlandsche Akademie van Wettenschoppen, Proceedings of the Section of Sciences, *10* (1947), 930–941.

Shell Development Company
Emeryville, California

(Received September 16, 1955)

A NOTE ON SALZER'S METHOD FOR SUMMING CERTAIN SLOWLY CONVERGENT SERIES

By P. Wynn

In a recent paper[1] Salzer has proposed a method of transforming certain slowly convergent series. He regards the partial sums S_n, where

$$S_n = \sum_{r=0}^{n} u_r$$

as functions of the variable $1/n$, and from the values of S_r for $r = n, n-1, \cdots n-m+1$, obtains the value of S_r when $1/r = 0$ by extrapolation, using Lagrange's interpolation formula. He tabulates the required Lagrangian interpolation coefficients for $m = 4, 7$ and 11; $n = 5, 10, 15$ and 20. Where the terms of the original series alternate in sign, the terms with odd and even suffices are grouped separately, and the two series thus formed are transformed as before. The method is clearly based on the approximation of the partial sums S_r, for r ranging from some n to ∞, by a polynomial in $1/r$, and were this approximation to be exact, as for example with the series $\sum_{n=2}^{\infty} 1/n(n-1)$, the method would succeed completely.

An alternative formulation of the method is now proposed whereby the transformation is applied directly to the terms of the series, and some indication of the error in the final result is provided. We apply the transformation to the partial sums S_n, S_{n+1}, S_{n+2}, $\cdots$ and obtain as an m^{th} order approximation to the sum

$$S \equiv \sum_{i=0}^{m-1} L_i^{(m-1)}(0) \cdot S_{n+i} \tag{1}$$

where

$$L_i^{(m-1)}(x) = \frac{(x-1/n)(x-1/(n+1)) \cdots (x-1/(n+i-1)) \cdot (x-1/(n+i+1)) \cdots (x-1/(n+m-1))}{(1/(n+i)-1/n)(1/(n+i)-1/(n+1)) \cdots (1/(n+i)-1/(n+i-1)) \cdot (1/n+i)-1/(n+i+1)) \cdots (1/(n+i)-1/(n+m-1))}$$

As Lagrange's interpolation formula and Newton's interpolation formula using divided differences are completely equivalent,[2] we may write (1) as

$$S \equiv S_n + (-1/n)\frac{S_{n+1}-S_n}{1/(n+1)-1/n} + (-1/n)(-1/(n+1)) \left(\frac{S_{n+2}-S_{n+1}}{1/(n+2)-1/(n+1)} - \frac{S_{n+1}-S_n}{1/(n+1)-1/n}\right)\frac{1}{1/(n+2)-1/n} + \cdots$$

$$\equiv S_n + (n+1)u_{n+1} + \frac{1}{2!}\frac{(n+2)}{(n+1)} \cdot \{(n+1)(n+2)u_{n+2} - n(n+1)u_{n+1}\} + \cdots$$

$$(2) \equiv S_n + \sum_{s=0}^{m-2} \frac{(n-1)!}{(n+s)!(s+1)!} \cdot v_0^s$$

$$(3) \equiv S_n + \sum_{s=0}^{m-2} t_s$$

where

$$(4) \qquad v_0^r = (n+r)(n+r+1)u_{n+r+1} \qquad r = 0, 1 \cdots, m-2$$

and

$$(5) \qquad v_r^{s+1} = (n+r)(n+r+s+2)(v_{r+1}^s - v_r^s) \qquad s = 0, 1 \cdots m-2,$$

It is here germane to point out that the transformation (2) may be prolonged at will by computing further terms in the sequences (4) and (5), and that we are in no way restricted by the necessity of using a predetermined number of points as in a formula of the form (1).

The remainder term in (1) and (2) is

$$R_{n,m} = (-1)^{m+1} \frac{(n-1)!}{(n+m-1)!} f(0, 1/n, 1/n+1, \cdots 1/n+m)$$

where $f(0, 1/n, 1/n+1 \cdots 1/n+m)$ is the $(m+1)^{\text{th}}$ order divided difference of S_n regarded as a function of $1/n$, for the $m+1$ arguments $0, 1/n, 1/n+1, \cdots 1/n+m$. We cannot of course, obtain this quantity directly since we do not know the value of S_n when $1/n = 0$, but some indication of the magnitude of $R_{n,m}$ is provided by the quantity $R'_{n,m}$, where

$$R'_{n,m} = (-1)^{m+1} \frac{(n-1)!}{(n+m-1)!} f(1/n+m+1, 1/n, 1/n+1, \cdots 1/n+m)$$
$$= t_{m-1},$$

t_{m-1} being the next term in the continuation of (2).

As an example of the use of the transformation (2) we may consider the series $\sum_{r=1}^{\infty} 1/r^2$ when $n = 4$, for which $S_4 = 1.423\ 611$. Table I shows successive terms t_s, partial sums S'_s, and remainders $R_{4,s}$ in (2).

TABLE I

s	t_s	S'_s	$R_{4,s}$
0	+0.2	1.623 611	+0.021 323
1	+0.02	1.643 611	+0.001 323
2	+0.001 296	1.644 907	+0.000 027
3	+0.000 031	1.644 938	−0.000 004
4	−0.000 003	1.644 935	−0.000 001

For series whose terms alternate in sign, we group the terms in pairs, and, if the terms of the series so derived form a monotonic sequence, proceed as before. For example, the series for $\log_e 2$, viz $\sum_{r=1}^{\infty} (-1)^r/r$ may be transformed as $\sum_{k=1}^{\infty} 1/2k(2k-1)$. Table II shows the derived sequences v_r^s for $n = 4$,

apocryphal figures having been successively deleted. Table III shows the quantities S'_s, $R_{4,s}$, and t_s, for this example

TABLE II

Table of v'_r for the series $\sum_{k=1}^{\infty} \frac{1}{2k(2k-1)}$, $n = 4$,

r	s					
	0	1	2	3	4	5
0	+0.22222 22222	+0.12121 212	+0.03263 41	−0.14919	−0.36	+1.0
1	0.22727 27273	0.12237 762	0.02797 19	0.1592	−0.33	
2	0.23076 92308	0.12307 692	0.02443 4	−0.1657		
3	0.23333 33333	0.12352 941	+0.02167 2			
4	0.23529 41177	+0.12383 901				
5	+0.23684 21053					

TABLE III

s	t_s	S'_s	$R_{4,s}$
0	+0.05555 556	0.69007 937	+0.00306 782
1	+0.00303 030	0.69310 967	+0.00003 751
2	+0.00004 533	0.69315 499	−0.00000 781
3	−0.00000 740	0.69314 759	−0.00000 041
4	−0.00000 045	0.69314 714	+0.00000 004
5	+0.00000 002	0.69314 716	+0.00000 002

Both Tables I and III show that the magnitudes of the successive remainders are in fact indicated by the succeeding terms in the transformed series (2). Table II is included to indicate the manner in which working figures are lost in the transformation, and to point out the need for the retention of extra working figures.

This note is published with the permission of the Director of the National Physical Laboratory.

REFERENCES

[1] H. E. Salzer, *A Simple Method for Summing Slowly Convergent Series*, Jour. Math. and Physics 33, pp. 356–359 (1955).

[2] L. M. Milne-Thomson, *The Calculus of Finite Differences*, Macmillan and Co. Ltd London, Chapter 1 p. 6.

National Physical Laboratory
Teddington, Middlesex, England

(Received July 19, 1955)

A NOTE ON THE SOLUTION OF A NON-LINEAR TRANSONIC FLOW EQUATION

By John T. Robacker and Albert D. Wheelon

Transition of the flow round a body with axial symmetry from subsonic to supersonic velocities is of considerable importance in aerodynamics. The slender body velocity potential ϕ in cylindrical coordinates satisfies a non-linear partial differential equation.[1]

$$(1 - M^2)\frac{\partial^2\phi}{\partial z^2} + \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right) = \alpha\frac{\partial\phi}{\partial z}\frac{\partial^2\phi}{\partial z^2}, \tag{A}$$

where z is the downstream coordinate and r the radial distance. As usual, M denotes the free stream Mach number and α is a (small) non-linear coupling constant. The non-linear portion of (A) is usually neglected. At sonic speeds, however, the first term vanishes (i.e., $M = 1$) and the linearized equation predicts infinite pressure coefficients on the body. The role of the non-linear term then is to supply a satisfactory transition from subsonic to supersonic conditions.

This note describes a technique for systematically generating particular solutions of the transonic equation:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right) = \alpha\frac{\partial\phi}{\partial z}\frac{\partial^2\phi}{\partial z^2}. \tag{B}$$

Let us equate both sides of this equation to an arbitrary combination of unknown functions of r alone and z alone. We shall call this function the "separator." Each of the resulting inhomogeneous differential equations may then be integrated partially with respect to its characteristic variable. The constants of integration, in each case, are functions of the other variable. Direct comparison of the two results then serves to identify the unknown functions.

The technique is best illustrated by an example. Let the separator be a function of z alone: $S = F(z)$.

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\phi}{\partial r}\right) = F(z) = \frac{\alpha}{2}\frac{\partial}{\partial z}\left(\frac{\partial\phi}{\partial z}\right)^2. \tag{C}$$

Integration on r yields

$$\phi(r, z) = \tfrac{1}{4}r^2F(z) + G(z)\ln(r) + H(z),$$

where $G(z)$ and $H(z)$ are the anticipated integration functions. The square of the first partial derivative of $\phi(r, z)$ with respect to z may be calculated from this expression and equated to the first (z) integral of the right-hand side of Eq. (C).

[1] R. Sauer, "General Characteristics of the Flow Through Nozzles at Near-Critical Speeds," NACA TM-1147.

$$\left(\frac{\partial\phi}{\partial z}\right)^2 = h(r) + \frac{2}{\alpha}\int F(z)\,dz$$

$$= \tfrac{1}{16}r^4[F'(z)]^2 + [G'(z)]^2\ln^2(r) + [H'(z)]^2$$

$$+ \tfrac{1}{2}r^2\ln(r)F'(z)G'(z) + \tfrac{1}{2}r^2F'(z)H'(z) + 2G'(z)H'(z)\ln(r)$$

Since r and z are not mixed in the first result, we infer that $F(z) = \eta$ = constant. The terms $\int F(z)\,dz$ and $[H'(z)]^2$ are the only functions depending on z alone, so that

$$H(z) = \frac{\alpha}{3\eta}\left[c + \frac{2\eta}{\alpha}z\right]^{\frac{3}{2}}.$$

In order that the last term above not be a function of both z and r, we must choose $G'(z) = 0$. When these identifications are all combined, one obtains,

$$\phi(r,z) = \frac{r^2}{4}\eta + \lambda\ln(r) + \frac{\alpha}{3\eta}\left[c + \frac{2\eta}{\alpha}z\right]^{\frac{3}{2}}. \tag{D}$$

One may also verify the following particular solutions of (B). The separator used in each example is indicated.

$S = f(r)$:

$$\phi(r,z) = \ln(r)\left[B + \frac{2\eta}{3}(c+z)^{\frac{3}{2}}\right] + A + \frac{2\lambda}{3}(c+z)^{\frac{3}{2}}$$

$$+ \frac{\alpha}{2}\int\frac{dr}{r}\int dr\,r[\lambda + \eta\ln(r)]^2 \tag{E}$$

$S = f(r) + F(z)$:

$$\phi(r,z) = \frac{\lambda}{4}zr^2 + \ln(r)[A + \eta z] + B + \frac{1}{2}\sqrt{\frac{\lambda}{\alpha}}z^2$$

$$+ \frac{\lambda\sqrt{\lambda\alpha}}{64}r^4 + \eta\sqrt{\lambda\alpha}\,r^2[\ln(r) - 1] \tag{F}$$

$S = 0$:

$$\phi(r,z) = (\eta z + B)\ln(r) + \lambda z + A. \tag{G}$$

All of these solutions (and their derivatives) exhibit divergent properties far from the body, so that one may not yet discuss the physical problem characterized by Eq. (A). An avenue of approach has been opened, however, which should serve this and similar problems.

THE RAMO-WOOLRIDGE CORP.
LOS ANGELES 45, CALIFORNIA

(Received August 12, 1955)

ELASTIC WAVES IN ANISOTROPIC MEDIA

By J. L. Synge*

Introduction. Against the background of 21-constant elastic theory, isotropy presents a highly degenerate case, and the study of waves in isotropic media throws little light on the algebraic structure of the theory of waves in anisotropic media. In this paper, some known results are set down in tensor notation, for the sake of convenience and compactness, and some new general results developed for waves in a layer of anisotropic material. In the limit in which the layer becomes a half-space, we have a generalization of Rayleigh waves, and it is found that in general (for a given frequency) the surface waves either do not exist at all, or are propagated in certain discrete directions with certain speeds. The paper ends with some calculations for a medium possessing transverse isotropy; in particular it is shown that the slowness surface (or, equivalently, the wave surface) possesses a spheroidal sheet.

1. General equations. Small latin suffixes take the values 1, 2, 3 with summation understood for a repeated suffix. Let x_p be rectangular cartesians. Let u_p be the displacement; then the strain tensor is

$$e_{pq} = \tfrac{1}{2}(u_{q,p} + u_{p,q}), \tag{1.1}$$

where the comma denotes partial differentiation ($u_{q,p} = \partial u_q/\partial x_p$). The stress tensor E_{pq} is connected with the strain tensor by the generalized Hooke's law:

$$E_{pq} = c_{pqrs}\, e_{rs}\,. \tag{1.2}$$

The elastic coefficients satisfy the symmetry conditions

$$c_{pqrs} = c_{qprs} = c_{pqsr} = c_{rspq}\,, \tag{1.3}$$

so that only 21 of them are independent. The strain energy function is

$$W = \tfrac{1}{2}E_{pq}\, e_{pq} = \tfrac{1}{2}c_{pqrs}\, e_{pq}\, e_{rs}\,, \tag{1.4}$$

and the sole condition imposed on the c's is the condition that W shall be positive-definite. Thus we write

$$W = \tfrac{1}{2}c_{pqrs}\, f_{pq}\, f_{rs} \geqq 0 \qquad \text{(p.d.)}, \tag{1.5}$$

where f_{mn} is any real symmetric tensor; this means that $W = 0$ implies $f_{mn} = 0$.

We shall consider only homogeneous media, so that the c's are constants. Since we are concerned with vibrations, the c's have their adiabatic values rather than their isothermal values.

* On leave of absence January to April, 1956, at the National Bureau of Standards, where this research was supported by the United States Air Force, through the Office of Scientific Research of the Research and Development Command.

The equations of motion are

$$\rho \frac{\partial^2 u_p}{\partial t^2} = E_{pq,q} + X_p, \tag{1.6}$$

where ρ is the density (assumed constant) and X_p is body force per unit volume.

Putting $X_p = 0$, and substituting from (1.1) and (1.2), we obtain the following equations:

$$\rho \frac{\partial^2 u_p}{\partial t^2} = c_{pqrs} u_{r,sq}. \tag{1.7}$$

These three partial differential equations of the second order have to be solved subject to appropriate boundary conditions. If the stress T_p is given across a surface with unit normal n_p, the boundary condition is

$$E_{pq} n_q = T_p. \tag{1.8}$$

2. Plane waves. We are concerned with plane elastic waves, that is, with solutions of (1.7) of the form

$$u_p = A_p \exp [i\omega(y_s x_s - t)], \tag{2.1}$$

where A_p, ω, y_s are seven constants, in general complex, the physical displacement being the real part of (2.1). However, we shall consider only disturbances which are periodic in t, and so ω is restricted to real values. We take $\omega > 0$ without loss of generality.

On substituting (2.1) in (1.7), we see that y_s must satisfy the characteristic equation

$$\Omega(y) = 0 \tag{2.2}$$

where

$$\Omega = \det M_{pr}, \tag{2.3}$$

$$M_{pr} = \rho\delta_{pr} - c_{pqrs} y_q y_s; \tag{2.4}$$

and when y_s has been so chosen, the amplitude vector A_p must satisfy

$$M_{pr} A_r = 0. \tag{2.5}$$

In the general case, the matrix M_{pr} is of rank 2, and so (2.5) determines the ratios of A_r. If B_r is a normalized solution of (2.5) (the particular normalization is of no importance), the general solution of (2.5) is of the form

$$A_r = fB_r \tag{2.6}$$

where f is arbitrary.

We can describe the whole class of plane waves as follows:

(i) Choose y_1, y_2, real or complex.

(ii) Solve the sextic (2.2) for y_3, choosing any one of the six solutions.

(iii) Determine B_r, the normalized solution of

$$M_{pr} B_r = 0. \tag{2.7}$$

(iv) Choose f arbitrarily, and obtain A_r from (2.6).

Thus any plane wave system of the type (2.1) depends on the choice of three arbitrary numbers, y_1, y_2, f, which may be real or complex.

The above procedure differs somewhat from the usual approach to plane waves in which the y's are taken to be real, and their ratios are assigned. Under such circumstances, (2.1) shows us plane waves propagated with speed

$$v = \frac{1}{\sqrt{y_s y_s}}, \tag{2.8}$$

in a direction with direction cosines n_s such that

$$y_s = \frac{n_s}{v}. \tag{2.9}$$

For this reason, we call y_s the *slowness vector*, following Hamilton. The characteristic equation (2.2) may now be written

$$\det | \rho v^2 \delta_{pr} - c_{pqrs} n_q n_s | = 0. \tag{2.10}$$

It follows from (1.5) that

$$c_{pqrs} X_p n_q X_r n_s \geqq 0 \qquad \text{(p.d.)} \tag{2.11}$$

(for n_p fixed and X_p arbitrary), and hence, by a well known result, the three roots v^2 of (2.10) are real and positive, so that there are three speeds of propagation in any assigned direction (say v', v'', v''')*.

The equation

$$\Omega(y) = 0 \tag{2.12}$$

defines a sextic surface in a space in which the y's are taken as rectangular cartesians. If the y's are complex, this is a complex surface. If the y's are real, as in (2.9), (2.12) gives us the *slowness surface*; it is such that each half-line drawn from the origin cuts the surface at three points, at distances $1/v'$, $1/v''$, $1/v'''$ from the origin.

The *wave surface* is the envelope of plane waves at time $t = 1$, the waves having passed through the origin at time $t = 0$. It is the polar reciprocal of the slowness surface. If Y_r are current coordinates on the wave surface, its equation is given in parametric form by

$$Y_r = \frac{\partial\Omega/\partial y_r}{y_s\, \partial\Omega/\partial y_s}, \qquad \Omega(y) = 0, \tag{2.13}$$

two of the y's being parameters.

The *velocity surface* is found by measuring off distances v', v'', v''', along each half-line drawn from the origin. It is therefore the inverse of the slowness surface.

Of these three surfaces, the slowness surface is the most fundamental mathematically, and it seems desirable to use Hamilton's descriptive name for it instead of the more usual "reciprocal wave surface."

* Cf. A. E. H. Love, Mathematical Theory of Elasticity (Cambridge Univ. Press, 1927; Dover Publications, New York, 1944), p. 298.

3. The problem of the excited layer. Consider a layer of anisotropic material occupying the region

$$-h < x_3 < 0. \tag{3.1}$$

The surface $x_3 = 0$ is free, and consequently we have on it the condition

$$E_{3q} = 0 \quad \text{for} \quad x_3 = 0. \tag{3.2}$$

The other surface $x_3 = -h$ is excited by application of stress waves

$$E_{3q} = C_q \exp[i\omega(y_1 x_1 + y_2 x_2 - t)], \tag{3.3}$$

where C_q, ω, y_1, y_2 are assigned constants; C_q may be complex, but ω, y_1, y_2 are real.*

We proceed to solve the equations of motion (1.7), with the above boundary conditions.

This is done in the following steps:

(i) Solve the equation

$$\Omega(y) = 0 \tag{3.4}$$

for y_2; as in (2.2), this is a sextic equation for y_3 with real coefficients, and we write

$$y_3^{(N)} \qquad (N = 1, 2, \cdots, 6) \tag{3.5}$$

for the roots, which we assume to be distinct. For notational reasons we define

$$y_1^{(N)} = y_1, \; y_2^{(N)} = y_2 \qquad (N = 1, 2, \cdots, 6). \tag{3.6}$$

Then the displacement

$$u_p = \sum_{N=1}^{6} A_p^{(N)} \exp[i\omega(y_s^{(N)} x_s - t)] \tag{3.7}$$

satisfies the equations of motion (1.7), provided the six amplitude vectors $A_n^{(N)}$ satisfy†

$$M_{pr}^{(N)} A_r^{(N)} = 0, \qquad (N = 1, 2, \cdots, 6) \tag{3.8}$$

where

$$M_{pr}^{(N)} = \rho\delta_{pr} - c_{pqrs} y_q^{(N)} y_s^{(N)}. \tag{3.9}$$

Let $B_r^{(N)}$ be normalized solutions of (3.8) [cf. (2.7)]; they are determined by y_1, y_2. Then, by (3.7), we have the following solutions of (1.7), $f^{(N)}$ being six arbitrary numbers:

$$u_p = \sum_{N=1}^{6} f^{(N)} B_p^{(N)} \exp[i\omega(y_m^{(N)} x_m^{(N)} - t)]. \tag{3.10}$$

The corresponding stress is

* We might say that the surface $x_3 = -h$ is *stroked*.

† Note that the summation convention does not operate for N, but only for small latin suffixes.

$$E_{pq} = i\omega \sum_{N=1}^{6} c_{pqrs} f^{(N)} B_r^{(N)} y_s^{(N)} \exp [i\omega(y_m^{(N)} x_m - t)]. \tag{3.11}$$

The boundary conditions at the free surface, as in (3.2), require

$$\sum_{N=1}^{6} c_{3qrs} f^{(N)} B_r^{(N)} y_s^{(N)} = 0, \tag{3.12}$$

and the boundary conditions at the excited face, as in (3.3), require

$$\sum_{N=1}^{6} c_{3qrs} f^{(N)} B_r^{(N)} y_s^{(N)} \exp (-i\omega h y_3^{(N)}) = -iC_q/\omega. \tag{3.13}$$

Solving the six equations (3.12), (3.13) for the six quantities

$$f^{(1)}, f^{(2)}, \cdots f^{(6)}, \tag{3.14}$$

and substituting in (3.10), we have the solution of the problem of the excited layer. The solution contains the following parameters

$$y_1, y_2, \omega, h, C_q, \tag{3.15}$$

of which the first four are real.

4. Resonance. Free vibrations.

Let us write (3.12), (3.13) in the form

$$\sum_{N=1}^{6} K_{MN} f^{(N)} = L_M, \qquad (M = 1, 2, \cdots, 6) \tag{4.1}$$

where

$$\begin{aligned}
K_{1N} &= c_{31rs} B_r^{(N)} y_s^{(N)}, \\
K_{2N} &= c_{32rs} B_r^{(N)} y_s^{(N)}, \\
K_{3N} &= c_{33rs} B_r^{(N)} y_s^{(N)}, \\
K_{4N} &= c_{31rs} B_r^{(N)} y_s^{(N)} \exp (-i\omega h y_3^{(N)}), \\
K_{5N} &= c_{32rs} B_r^{(N)} y_s^{(N)} \exp (-i\omega h y_3^{(N)}), \\
K_{6N} &= c_{33rs} B_r^{(N)} y_s^{(N)} \exp (-i\omega h y_3^{(N)}), \\
&\qquad\qquad (N = 1, 2, \cdots, 6)
\end{aligned} \tag{4.2}$$

and

$$L_1 = L_2 = L_3 = 0, \qquad i\omega L_4 = C_1, \qquad i\omega L_5 = C_2, \qquad i\omega L_6 = C_3. \tag{4.3}$$

Let

$$\Delta(y_1, y_2, \omega, h) = \det K_{MN}, \tag{4.4}$$

a 6×6 determinant which is a function of the four arguments shown.

If we hold C_q fixed and change y_1, y_2, ω, h so that Δ tends to zero, then the displacement (3.10) tends to infinity, and resonance occurs. Thus the condition for resonance is

$$\Delta(y_1, y_2, \omega, h) = 0. \tag{4.5}$$

This is a complex equation, equivalent to two real equations, say,

$$\Delta_1(y_1, y_2, \omega, h) = 0, \qquad \Delta_2(y_1, y_2, \omega, h) = 0. \tag{4.6}$$

These are two surfaces in a 4-space in which y_1, y_2, ω, h are coordinates, and these surfaces may intersect or they may not.

If we fix ω, h, then (4.6) are two curves in the plane of y_1, y_2. The vector (y_1, y_2) is the slowness vector of the surface waves. If the two curves in question do not intersect, resonance is impossible; if they intersect, the intersections will in general form a discrete set of points, indicating a discrete set of resonant slownesses.

Free vibrations are of course closely connected with resonance. Let us remove the excitation by putting $C_q = 0$. Then we have in (4.1) a set of six homogeneous equations, which are consistent if, and only if, the determinantal equation (4.5) is satisfied. If ω and h are assigned, there are (apart from cases of degeneracy) either (i) no possible free vibrations, or (ii) a discrete set of free vibrations, each with a certain surface slowness vector (y_1, y_2).

If we put $y_1 = y_2 = 0$, then the six roots $y_3^{(N)}$ are real and of the form

$$y_3^{(N)} = \pm\alpha', \qquad \pm\alpha'', \qquad \pm\alpha'''. \tag{4.7}$$

The quantities $B_r^{(N)}$, satisfying (3.8) are then real, and it is easy to see that the determinant Δ of (4.4) is a pure imaginary. In this case we have one condition (and not two) for resonance or free vibration.

5. Rayleigh waves. Let us now consider Rayleigh waves in an anistropic half-space $x_3 < 0$ by proceeding to the limit $h \to \infty$ in the preceding theory.

But before doing this, we must consider an algebraic question concerning the slowness equation

$$\Omega(y) = 0, \tag{5.1}$$

which we now regard as a sextic equation for y_3, real values being assigned to y_1, y_2.

If we put $y_1 = y_2 = 0$, then (5.1) gives six real values to y_3, of the form

$$y_3 = \pm\alpha', \qquad \pm\alpha'', \qquad \pm\alpha''', \tag{5.2}$$

where α', α'', α''' are the reciprocals of the three speeds of propagation in the direction of the x_3-axis.

In a plane P in which y_1, y_2 are rectangular Cartesians, draw a half-line L out from the origin, and lead a point (y_1, y_2) along this half-line. The roots of (5.1) must change continuously, and therefore for some segment OA of L the six values of y_3 are real (but not necessarily symmetrically distributed as in (5.2)).

From the positive definite condition (1.5) it follows that (5.1) represents a bounded surface, and hence if we go to infinity on L we shall ultimately find that all six values of y_3 are complex; since the coefficients in (5.1) are real, such complex roots occur in conjugate pairs $(\alpha \pm i\beta)$.

If we reverse the signs of y_1, y_2, we reverse the signs of roots y_3. Hence we

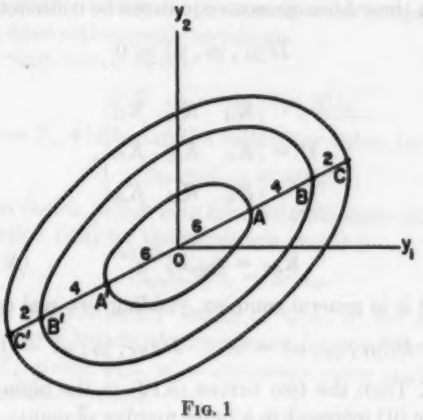


FIG. 1

can prepare a diagram as Fig. 1, which shows the plane of y_1, y_2 divided into four regions. $C'B'A'$ $OABC$ is a typical straight line through the origin. For (y_1, y_2) in the segment $A'A$, there are six real roots y_3; in $B'A'$ and in BA there are four; in $C'B'$ and in BC there are two. Further out, there are none. As the typical line swings around, the segments trace out regions as shown. Note that

$$C'B' = BC, \qquad B'A' = AB, \qquad A'O = OA. \tag{5.3}$$

Consider now a displacement of the form

$$u_p = A_p \exp [i\omega(y_s x_s - t)]. \tag{5.4}$$

If $y_3 = \alpha + i\beta$, this reads

$$u_p = A_p \exp (-\omega\beta x_3) \exp [i\omega(y_1 x_1 + y_2 x_2 + \alpha x_3 - t)] \tag{5.5}$$

If we demand that this tends to zero as $x_3 \to -\infty$, we must have β negative. This means that, in the problem of surface waves, real roots of (5.1) are of no use to us, and, of the complex roots, we can use only those with negative imaginary parts. Therefore we must confine our attention to values of y_1, y_2 outside the outer oval in Fig. 1. For any such choice of real y_1, y_2, we have six values of y_3 of the form

$$y_3 = \alpha' \pm i\beta', \qquad \alpha'' \pm i\beta'', \qquad \alpha''' \pm i\beta'''. \tag{5.6}$$

Let $y_3^{(N)} (N = 1, 2, 3)$ be the three of these with negative imaginary parts. Let $B_r^{(N)} (N = 1, 2, 3)$ be corresponding normalized solutions of (3.8). Let $f^{(N)} (N = 1, 2, 3)$ be any three numbers. Then the disturbance

$$u_p = \sum_{N=1}^{3} f^{(N)} B_p^{(N)} \exp [i\omega(y_s^{(N)} x_s - t)] \tag{5.7}$$

satisfies the equations of motion and vanishes exponentially at $x_3 = -\infty$.

This disturbance will make $x_3 = 0$ a free surface provided $f^{(N)}$ satisfy

$$\sum_{N=1}^{3} c_{3qrs} f^{(N)} B_r^{(N)} y_s^{(N)} = 0. \tag{5.8}$$

The condition that these homogeneous equations be consistent is

$$D(y_1, y_2, \omega) = 0 \tag{5.9}$$

where

$$D = \begin{vmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{vmatrix}, \tag{5.10}$$

with

$$K_{qN} = c_{2qrs} B_r^{(N)} y_s^{(N)} \qquad (N = 1, 2, 3). \tag{5.11}$$

The equation (5.9) is in general complex, yielding two real equations

$$D_1(y_1, y_2, \omega) = 0, \qquad D_2(y_1, y_2, \omega) = 0. \tag{5.12}$$

Let ω be assigned. Then the two curves (5.12) in the plane of y_1, y_2 either (i) do not intersect, or (ii) intersect in a finite number of points, if we leave degenerate cases out of account.

The disturbance (5.7), which dies out exponentially as we travel downward in the medium, is the generalization of Rayleigh waves to the general anisotropic case. We conclude that, for given circular frequency ω, such waves either (i) do not exist, or (ii) travel only in certain directions in the free surface, with certain corresponding speeds. The familiar Rayleigh waves are degenerate, the locus of possible points (y_1, y_2) being a circle, and not a number of isolated points.

A further difference between the ordinary Rayleigh waves and those considered above is that the latter exhibit a sinusoidal variation with depth. This is due to the fact that in (5.6), α', α'', α''' are in general different from zero,* whereas in the ordinary Rayleigh waves y_3 is a pure imaginary, giving a purely exponential decay with depth. One may say that the ordinary Rayleigh waves are propagated parallel to the free surface, with amplitudes decreasing exponentially with depth, whereas the directions of propagation of the generalized Rayleigh waves are inclined to the free surface.

6. Transverse isotropy. Here follow some remarks about media possessing transverse isotropy. For such a medium, all directions orthogonal to some unit vector J_r are elastically equivalent, and we have

$$\begin{aligned} c_{pqrs} = {} & \lambda\delta_{pq}\delta_{rs} + \mu(\delta_{pr}\delta_{qs} + \delta_{ps}\delta_{qr}) + \alpha(\delta_{pq}J_rJ_s + \delta_{rs}J_pJ_q) \\ & + \beta(\delta_{pr}J_qJ_s + \delta_{qr}J_pJ_s + \delta_{ps}J_qJ_r + \delta_{qs}J_pJ_r) + \gamma J_pJ_qJ_rJ_s \end{aligned} \tag{6.1}$$

where λ, μ, α, β, γ are elastic constants. We get an isotropic medium on putting $\alpha = \beta = \gamma = 0$.

Let us investigate the slowness surface,

$$\Omega(y) = \det M_{pr} = 0, \qquad M_{pr} = \rho\delta_{pr} - c_{pqrs}y_qy_s\,. \tag{6.2}$$

* See §6 for the case of transverse isotropy.

Equivalently we may investigate the speeds v for propagation in a direction n_p. This may be done rather neatly as follows.

Let X_p be any unit vector. Write

$$Q(X, n) = c_{pqrs}X_p n_q X_r n_s . \tag{6.3}$$

Then those vectors X_p which give Q a stationary value, for assigned n_p, satisfy

$$c_{pqrs}n_q X_r n_s = \phi X_p \tag{6.4}$$

where ϕ is a scalar factor, which is in fact the stationary value of Q. But, in view of (2.9), the equation (2.5) for the amplitude vector is

$$c_{pqrs}n_q A_r n_s = \rho v^2 A_p . \tag{6.5}$$

We conclude that the stationary values of $Q(X, n)$ are $\rho v'^2$, $\rho v''^2$, $\rho v'''^2$ where v', v'', v''' are the three speeds of propagation for the direction n_p, and further that the vector X_p giving $Q(X, n)$ a stationary value is in the direction of the corresponding amplitude vector.

In the case of transverse isotropy we have

$$\begin{aligned} Q(X, n) = \mu + \beta(Jn)^2 + (\lambda + \mu)(Xn)^2 + [\beta + \gamma(Jn)^2](XJ)^2 \\ + 2(\alpha + \beta)(Jn)(XJ)(Xn), \end{aligned} \tag{6.6}$$

where $(UV) = U_r V_r$, the scalar product. One stationary value is given by taking X_p orthogonal to J_p and to n_p; hence we get the speed v' given by

$$\rho v'^2 = \mu + \beta(Jn)^2. \tag{6.7}$$

Thus one sheet of the slowness surface $\Omega = 0$ is a spheroid, its equation in polar coordinates being

$$\mu r^2 + \beta r^2 \cos^2\theta = \rho, \tag{6.8}$$

where $r^2 = y_1^2 + y_2^2 + y_3^2$ and θ is the angle between n_p and the fixed vector J_p.

The rest of the slowness surface is a quartic surface. To investigate it, we put

$$X_p = AJ_p + Bn_p , \tag{6.9}$$

where

$$A^2 + B^2 + 2AB(Jn) = 1, \tag{6.10}$$

since X_p is a unit vector. We have

$$(Xn) = A(Jn) + B, \qquad (XJ) = A + B(Jn), \tag{6.11}$$

and substitution in (6.6) gives

$$Q(X, n) = \mu + \beta(Jn)^2 + KA^2 + 2LAB + MB^2, \tag{6.12}$$

where

$$\begin{aligned} K &= (\lambda + \mu)(Jn)^2 + [\beta + \gamma(Jn)^2] + 2(\alpha + \beta)(Jn)^2, \\ L &= (\lambda + \mu)(Jn) + [\beta + \gamma(Jn)^2](Jn) + (\alpha + \beta)(Jn)[1 + (Jn)^2], \\ M &= (\lambda + \mu) + [\beta + \gamma(Jn)^2](Jn)^2 + 2(\alpha + \beta)(Jn)^2. \end{aligned} \tag{6.13}$$

To give $Q(X, n)$ a stationary value, we require

$$KA + LB = \psi[A + B(Jn)], \qquad LA + MB = \psi[A(Jn) + B], \tag{6.14}$$

where ψ is a Lagrange multiplier; actually

$$\psi = \rho v^2 - \mu - \beta(Jn)^2. \tag{6.15}$$

Thus the two other speeds of propagation are given by

$$\begin{vmatrix} \psi - K & \psi(Jn) - L \\ \psi(Jn) - L & \psi - M \end{vmatrix} = 0, \tag{6.16}$$

or

$$\psi^2 \sin^2\theta - \psi(K + M - 2L\cos\theta) + KM - L^2 = 0, \tag{6.17}$$

where

$$\cos\theta = (Jn). \tag{6.18}$$

In polar coordinates, the equation of the quartic sheet of the slowness surface has the equation (6.17), with

$$\psi = \frac{\rho}{r^2} - \mu - \beta\cos^2\theta. \tag{6.19}$$

The displacement vector A_p is perpendicular to the meridian plane in the case of the spheroidal sheet (6.8). In the case of the quartic sheet, it is given by (6.9), after we substitute the appropriate values of A, B, given by (6.14).

If we degenerate to the case of isotropy by putting $\alpha = \beta = \gamma = 0$, the spheroidal sheet (6.8) becomes the sphere

$$r = \frac{1}{v'} = \sqrt{\frac{\rho}{\mu}}, \tag{6.20}$$

corresponding to an S-wave. We have

$$\begin{aligned} &K = (\lambda + \mu)\cos^2\theta, \qquad L = (\lambda + \mu)\cos\theta, \qquad M = \lambda + \mu, \\ &KM - L^2 = 0, \qquad K + M - 2L\cos\theta = (\lambda + \mu)\sin^2\theta, \end{aligned} \tag{6.21}$$

so that (6.17) gives

$$\psi = 0 \text{ or } \lambda + \mu, \tag{6.22}$$

and the quartic sheet (6.19) becomes the two spheres

$$r = \frac{1}{v''} = \sqrt{\frac{\rho}{\mu}}, \qquad r = \frac{1}{v'''} = \sqrt{\frac{\rho}{\lambda + 2\mu}}, \tag{6.23}$$

corresponding respectively to an S-wave and to a P-wave.

The case of a slight variation from isotropy can be investigated easily, second-order terms in α, β, γ being neglected. Naturally, the slowness surface has two

sheets (one spheroidal) which are close to a sphere of radius $\sqrt{\rho/\mu}$, and the third sheet close to a smaller sphere of radius $\sqrt{\rho/(\lambda + 2\mu)}$.

It is easy to verify, in the case of the spheroidal sheet (6.8), the statement of §5 that the generalized Rayleigh waves have a sinusoidal variation along the downward vertical. This is the same thing as saying that y_3 has a non-vanishing real part. We know that

$$\Omega(y) = [\rho - \mu y_s y_s - \beta(J_s y_s)^2]F(y) \tag{6.24}$$

where $F(y)$ is a quartic factor. Thus two of the six values of y_3, corresponding to assigned real values of y_1, y_2, are given by solving the quadratic equation

$$\begin{aligned}(\mu + \beta J_3^2)y_3^2 + 2\beta y_3 J_3(J_1 y_1 + J_2 y_2) \\ + \mu(y_1^2 + y_2^2) + \beta(J_1 y_1 + J_2 y_2)^2 - \rho = 0.\end{aligned} \tag{6.25}$$

The roots are

$$y_3 = \frac{1}{\mu + \beta J_3^2}\{-\beta J_3(J_1 y_1 + J_2 y_2) \pm iR\}, \tag{6.26}$$

where

$$R^2 = (\mu + \beta J_3^2)[\mu(y_1^2 + y_2^2) - \rho] + \mu\beta(J_1 y_1 + J_2 y_2)^2. \tag{6.27}$$

In order that Rayleigh waves may exist, it is necessary that R^2 be positive. Then the real part of y_3 is

$$\operatorname{Re} y_3 = -\frac{\beta J_3(J_1 y_1 + J_2 y_2)}{\mu + \beta J_3^2}. \tag{6.28}$$

This vanishes only under one of the following three conditions:

(i) $\beta = 0$ (as is the case in isotropy);

(ii) $J_3 = 0$ (axis of symmetry parallel to the free surface);

(iii) $J_1 y_1 + J_2 y_2 = 0$.

This last condition is satisfied when the axis of symmetry is perpendicular to the free surface, or, more generally, when the surface slowness vector $(y_1 y_2)$ is perpendicular to the projection of the axis of symmetry on the free surface.

In recent papers, Musgrave [2] has discussed body waves in a medium with transverse isotropy (which is the same as hexagonal symmetry, as far as elastic properties are concerned). He follows the method described by Kelvin in his Baltimore Lectures, and the calculations look very different from those of the present paper. Musgrave appears to have overlooked the interesting factorization (6.24) corresponding to a spheroidal sheet of the slowness surface. But it is not hard to get this result out of his work, for his characteristic cubic factorizes as follows:

$$\begin{aligned}H^3 - H^2\{n^2 h + (1 - n^2)(a + \tfrac{1}{2}c)\} \\ + H(1 - n^2)\{(1 - n^2)\tfrac{1}{2}ac + n^2[h(a + \tfrac{1}{2}c) - d^2]\} \\ - n^2(1 - n^2)^2\tfrac{1}{2}c(ah - d^2) = \{H - \tfrac{1}{2}c(1 - n^2)\} \\ \{H^2 - H[n^2 h + (1 - n^2)a] + n^2(1 - n^2)(ah - d^2)\}.\end{aligned} \tag{6.29}$$

The vanishing of the first factor gives the spheroidal sheet.

To connect the above work more closely with Musgrave's, we may consider the intersections of the slowness surface with the axis of symmetry J_r, or, equivalently, the speeds of propagation in the direction $n_r = J_r$. In this case, we put $(Jn) = 1$, $(Xn) = (XJ)$ in (6.6), and obtain

$$Q(X, n) = \mu + \beta + (\lambda + \mu + 2\alpha + 3\beta + \gamma)(XJ)^2. \tag{6.30}$$

The stationary values are given by $(XJ) = 0$, $(XJ) = 1$, and hence we get, for the three velocities

$$\begin{aligned} \rho v'^2 &= \mu + \beta, \qquad \rho v''^2 = \mu + \beta, \\ \rho v'''^2 &= \lambda + 2\mu + 2\alpha + 4\beta + \gamma, \end{aligned} \tag{6.31}$$

where v' comes from the spheroidal sheet, and v'', v''' from the quartic sheet. In fact, the spheroidal sheet touches the quartic sheet on the axis of symmetry, as is shown in Musgrave's diagrams, in which T_1 is the spheroidal sheet in his Figs. 1b and 2b. The polar reciprocal of this spheroidal sheet appears as T_1 in his Figs. 1c and 2c.

Stoneley [1] has discussed Rayleigh waves in the case of transverse isotropy with the axis of symmetry perpendicular to the free surface, and also surface waves in a cubic crystal [3]. Conical refraction in a cubic crystal has been treated by de Klerk and Musgrave [4].

Added in proof: Since the above paper was written, I have investigated the directions of the flow of energy for plane waves in a layer of anisotropic elastic material; see Proc. Roy. Irish Acad. A 58, 13–21 (1956).

REFERENCES

1. R. Stoneley, Monthly Notices Roy. Ast. Soc., Geophysical Supplement *5*, 343–353 (1949).
2. M. J. P. Musgrave, Proc. Roy. Soc. London A 226, 339–366 (1954).
3. R. Stoneley, Proc. Roy. Soc. London A 232, 447–458 (1955).
4. J. de Klerk and M. J. P. Musgrave, Proc. Phys. Soc. B 68, 81–88 (1955).

Dublin Institute for Advanced Studies

(Received April 9, 1956)

DEFORMATION ENERGY AND THE STRESS-STRAIN RELATIONS FOR ISOTROPIC MATERIALS

By T. Y. Thomas

1. Introduction. The rate at which work is done on a moving volume $V(t)$ of a material medium, subject to the usual laws of dynamics, by the stresses acting over its surface is given by

$$W = \frac{d}{dt}\int_{V(t)} \frac{1}{2}\rho v_\alpha v_\alpha \, dV + \int_{V(t)} \sigma_{\alpha\beta}\epsilon_{\alpha\beta} \, dV, \tag{1.1}$$

where $\sigma_{\alpha\beta}$ and $\epsilon_{\alpha\beta}$ are the components of the symmetric stress tensor σ and the rate of strain tensor ϵ respectively, v_α the components of velocity and ρ the density of the material. Equation (1.1) and following equations are referred to rectangular Cartesian coordinates used as Eulerian variables. The first term in the right member of (1.1) is the rate of change of the *kinetic energy* and the second term represents the rate of change of the *deformation energy* of the moving volume $V(t)$. Denoting this latter energy per unit mass by E we can write[1]

$$\frac{d}{dt}\int_{V(t)} \rho E \, dV \int_{V(t)} \rho \frac{dE}{dt} \, dV = \int_{V(t)} \sigma_{\alpha\beta}\epsilon_{\alpha\beta} \, dV, \tag{1.2}$$

and from this equation it follows that

$$\rho \frac{dE}{dt} = \sigma_{\alpha\beta}\epsilon_{\alpha\beta}, \tag{1.3}$$

since the volume V can be selected arbitrarily at any initial time.

Under what conditions will the energy E be a scalar invariant of the tensors σ and ϵ and the density ρ of an isotropic medium? This question is investigated in the following paper[2]. The answer depends on the specific character of the stress-strain relations which we shall here assume to have the general form

$$\frac{D\sigma_{\alpha\beta}}{Dt} = f_{\alpha\beta}(\rho, \sigma, \epsilon), \tag{1.4}$$

where the f's are the components of a tensor invariant under proper orthogonal transformations relating moving rectangular coordinate systems x and the quantities $D\sigma_{\alpha\beta}/Dt$ are the components of the *absolute time derivative*[3] of the stress tensor. In addition there are the general dynamical relations

[1] An expository treatment of the basic assumptions and the mathematical analysis underlying the derivation of the equations (1.1) and (1.2) has been given by T. Y. Thomas, *The fundamental hydrodynamical equations and shock conditions for gases*, Math. Mag., **22**, 1949, pp. 169–189.

[2] Prepared for the Applied Mathematics Branch, Mechanics Division, Naval Research Laboratory, Washington, D. C.

[3] T. Y. Thomas, *On the structure of the stress-strain relations*, Proc. Nat. Acad. Sci., **41**, 1955, pp. 716–720; and *Kinematically preferred coordinate systems*, *ibid.*, **41**, 1955, pp. 762–770. See also, C. Truesdell, *The simplest rate theory of pure elasticity*, Comm. Pure and Appl. Math., **8**, 1955, pp. 123–132.

$$\frac{d\rho}{dt} + \rho v_{\alpha,\alpha} = 0; \qquad \sigma_{\alpha\beta,\beta} = \rho \frac{dv_\alpha}{dt} \tag{1.5}$$

of which the first will also enter into the discussion of the problem. These latter relations involve the ordinary or total derivative with respect to the time t which we distinguish from the absolute time derivative in (1.4). The comma in (1.5) and following relations stands for partial differentiation since rectangular coordinates are employed exclusively.

We shall restrict the functions $f_{\alpha\beta}$ to be polynomials in the components $\sigma_{\alpha\beta}$ and $\epsilon_{\alpha\beta}$ with coefficients depending on the density ρ. To avoid excessive mathematical details we shall furthermore impose the condition that the polynomials $f_{\alpha\beta}$ are *linear* in the $\epsilon_{\alpha\beta}$ and at most *quadratic* in the $\sigma_{\alpha\beta}$. The precise tensorial form assumed for these functions f is given by the equation (6.1) in §6. The resulting equations (1.4) have been called *preferred* for purpose of reference[4]. It may be noted however that in spite of the restrictions underlying the preferred system (1.4) such systems are sufficiently general to include the Prandtl-Reuss equations[5] for the treatment of plastic flow, which may arise for large stresses, as well as to permit an approximation, for small stresses, to the results of the classical elasticity theory. It would appear in fact that the preferred systems are not too general for the requirement of application but are nevertheless of sufficient generality to provide an acceptable description of the dynamical behavior of certain important material bodies.

In §2 we have derived a modification of the differential equation (1.3) for the energy function E and in §6 we have given the set of differential equations into which this latter equation decomposes for the case of the preferred system of stress-strain relations. As indicated in §6 there appears to be a natural class of solutions of the differential equations for the determination of the deformation energy E, based on the structure of these equations, when the energy E does not depend on the density ρ of the medium. Special preferred stress-strain relations and the associated energy functions E of this type have been constructed in §§8 to 11. One of these functions E is observed to be identical with the well known expression for the deformation energy in the classical elasticity theory and another to approximate the classical expression for small stress.

In §13 the case is treated for which the deformation energy E depends on the density ρ and the corresponding results are obtained.

2. General differential equation for the energy function. If the deformation energy E can be expressed as a (differentiable) scalar invariant of the tensors σ and ϵ and the density ρ we can write

$$\frac{dE}{dt} = \frac{DE}{Dt} = \frac{\partial E}{\partial \rho}\frac{d\rho}{dt} + \frac{\partial E}{\partial \sigma_{\alpha\beta}}\frac{D\sigma_{\alpha\beta}}{Dt} + \frac{\partial E}{\partial \epsilon_{\alpha\beta}}\frac{D\epsilon_{\alpha\beta}}{Dt},$$

[4] The preferred system would be a special case of the hypo-elastic system of grade two as defined by Truesdell except for the fact that the $f_{\alpha\beta}$ are not assumed to be homogeneous in the components $\epsilon_{\alpha\beta}$. See, C. Truesdell, *Hypo-elasticity*, J. Rational Mech. Anal., **4**, 1955, pp. 83–133.

[5] See, for example, W. Prager and P. G. Hodge, *The Theory of Perfectly Plastic Solids*, New York, John Wiley & Sons, 1951, p. 29.

on account of the equality of the total and absolute time derivatives of a scalar. Hence from (1.4) and (1.5) the equation (1.3) becomes

$$-\rho\epsilon_{\alpha\alpha}\frac{\partial E}{\partial \rho}+f_{\alpha\beta}\frac{\partial E}{\partial \sigma_{\alpha\beta}}+\frac{\partial E}{\partial \epsilon_{\alpha\beta}}\frac{D\epsilon_{\alpha\beta}}{Dt}=\frac{1}{\rho}\sigma_{\alpha\beta}\epsilon_{\alpha\beta}. \tag{2.1}$$

Now the equation (2.1) must be satisfied identically in the quantities ρ, $\sigma_{\alpha\beta}$, $\epsilon_{\alpha\beta}$, and their partial derivatives considered as independent variables in order for the invariant E to exist in the proper sense[6]. But from the definition of the absolute time derivative we have

$$\frac{D\epsilon_{\alpha\beta}}{Dt}=\frac{\partial\epsilon_{\alpha\beta}}{\partial t}+\epsilon_{\alpha\beta,\gamma}v_\gamma+\epsilon_{\mu\beta}\phi_{\mu\alpha}+\epsilon_{\alpha\mu}\phi_{\mu\beta},$$

where the ϕ's are the components of the (non-tensor) curl of the velocity.[3] Hence we immediately see that the partial derivatives $\partial E/\epsilon_{\alpha\beta}$ in (2.1) must vanish identically. *In other words, the invariant E can depend at most on the stress tensor σ and the density ρ.*

Let us now put

$$\theta=\sigma_{\alpha\alpha};\qquad \xi=\sigma_{\alpha\beta}\sigma_{\alpha\beta};\qquad \zeta=\sigma_{\alpha\beta}\sigma_{\alpha\gamma}\sigma_{\beta\gamma}. \tag{2.2}$$

Then it is well known that the scalar invariant $E(\rho, \sigma)$ can be expressed as a function of the quantities ρ, θ, ξ, and ζ. But by differentiation of the relations (2.2) we have

$$\frac{\partial\theta}{\partial\sigma_{\alpha\beta}}=\delta_{\alpha\beta};\qquad \frac{\partial\xi}{\partial\sigma_{\alpha\beta}}=2\sigma_{\alpha\beta};\qquad \frac{\partial\zeta}{\partial\sigma_{\alpha\beta}}=3\sigma_{\alpha i}\sigma_{i\beta} \tag{2.3}$$

when no regard is paid to the symmetric character of the components $\sigma_{\alpha\beta}$ in the above expressions (2.2) preceding the process of differentiation. Hence we can write.

$$\frac{\partial E}{\partial\sigma_{\alpha\beta}}=\frac{\partial E}{\partial\theta}\delta_{\alpha\beta}+2\frac{\partial E}{\partial\xi}\sigma_{\alpha\beta}+3\frac{\partial E}{\partial\zeta}\sigma_{\alpha i}\sigma_{i\beta}.$$

Substituting this expression for the derivatives of E into (2.1) this relation can be given the following form

$$-\rho\epsilon_{\alpha\alpha}\frac{\partial E}{\partial\rho}+f_{\alpha\alpha}\frac{\partial E}{\partial\theta}+2\sigma_{\alpha\beta}f_{\alpha\beta}\frac{\partial E}{\partial\xi}+3\sigma_{\alpha\beta}\sigma_{\alpha\gamma}f_{\beta\gamma}\frac{\partial E}{\partial\zeta}=\frac{1}{\rho}\sigma_{\alpha\beta}\epsilon_{\alpha\beta}.$$

We shall later find it convenient to employ the "trace" notation on the basis of which we may, for example, write

$$\theta=\operatorname{tr}\mathfrak{d};\qquad \xi=\operatorname{tr}\mathfrak{d}^2;\qquad \zeta=\operatorname{tr}\mathfrak{d}^3.$$

Using this notation the above differential equation for E becomes

$$-\rho\operatorname{tr}\mathfrak{e}\frac{\partial E}{\partial\rho}+\operatorname{tr}\mathfrak{f}\frac{\partial E}{\partial\theta}+2\operatorname{tr}\mathfrak{d}\mathfrak{f}\frac{\partial E}{\partial\xi}+3\operatorname{tr}\mathfrak{d}^2\mathfrak{f}\frac{\partial E}{\partial\zeta}=\frac{1}{\rho}\operatorname{tr}\mathfrak{d}\mathfrak{e}. \tag{2.4}$$

[6] Without this requirement the differential equation (2.1) would admit, in general, only a solution E dependent on the coordinates x and the time t and this would be inconsistent with the strict condition that E is a scalar invariant as postulated.

It will be shown in the following sections that the differential equation (2.4) must decompose into a set of differential equations involving only the function E and the independent variables ρ, θ, ξ, and ζ. A scalar invariant $E(\rho, \sigma)$ will therefore represent the deformation energy of the material medium governed by (1.4) and (1.5) if, and only if, it satisfies this set of differential equations.

3. Stress-strain matrix. The matrix f whose elements are the symmetric components $f_{\alpha\beta}$ wil be called the *stress-strain matrix*. We assume the quantities $f_{\alpha\beta}$ to be *polynomials* in the components $\sigma_{\alpha\beta}$ and $\epsilon_{\alpha\beta}$ of the stress and rate of strain tensors. Then according to Rivlin[7] the matrix $\mathbf{f}$ must be expressible in the form

$$\begin{aligned}\mathbf{f} = a_0\mathbf{I} + a_1\mathfrak{d} + a_2\mathfrak{d}^2 + a_3\mathfrak{e} + a_4(\mathfrak{d}\mathfrak{e} + \mathfrak{e}\mathfrak{d}) + a_5(\mathfrak{d}^2\mathfrak{e} + \mathfrak{e}\mathfrak{d}^2) \\ + a_6\mathfrak{e}^2 + a_7(\mathfrak{d}\mathfrak{e}^2 + \mathfrak{e}^2\mathfrak{d}) + a_8(\mathfrak{d}^2\mathfrak{e}^2 + \mathfrak{e}^2\mathfrak{d}^2),\end{aligned} \tag{3.1}$$

where $\mathbf{I}$ denotes the unit matrix and the symbols $\mathfrak{d}$ and $\mathfrak{e}$ denote the stress and rate of strain matrices; moreover, the coefficients $a_0, \cdots, a_8$ are polynomials in the following quantities

$$\begin{cases}\operatorname{tr}\mathfrak{d}; & \operatorname{tr}\mathfrak{e}; & \operatorname{tr}\mathfrak{d}^2; & \operatorname{tr}\mathfrak{e}^2; & \operatorname{tr}\mathfrak{d}^3; & \operatorname{tr}\mathfrak{e}^3, \\ & \operatorname{tr}\sigma\mathfrak{e}; & \operatorname{tr}\mathfrak{d}^2\mathfrak{e}; & \operatorname{tr}\mathfrak{d}\mathfrak{e}^2; & \operatorname{tr}\mathfrak{d}^2\mathfrak{e}^2. & \end{cases} \tag{3.2}$$

Obviously the coefficients of the polynomials $a_0, \cdots, a_8$ will be functions of ρ since we have assumed the matrix $\mathbf{f}$ to depend on the density of the material (§1).

We now make the fundamental assumption that the relation (3.1) *is linear in the quantities* $\epsilon_{\alpha\beta}$. Then (3.1) becomes

$$\mathbf{f} = a_0\mathbf{I} + a_1\mathfrak{d} + a_2\mathfrak{d}^2 + a_3\mathfrak{e} + a_4(\mathfrak{d}\mathfrak{e} + \mathfrak{e}\mathfrak{d}) + a_5(\mathfrak{d}^2\mathfrak{e} + \mathfrak{e}\mathfrak{d}^2), \tag{3.3}$$

and the coefficients a_0, a_1 and a_2 are given by equations of the form

$$\begin{cases} a_0 = p_0 + q_0 \operatorname{tr}\mathfrak{e} + r_0 \operatorname{tr}\mathfrak{d}\mathfrak{e} + s_0 \operatorname{tr}\mathfrak{d}^2\mathfrak{e}, \\ a_1 = p_1 + q_1 \operatorname{tr}\mathfrak{e} + r_1 \operatorname{tr}\mathfrak{d}\mathfrak{e} + s_1 \operatorname{tr}\mathfrak{d}^2\mathfrak{e}, \\ a_2 = p_2 + q_2 \operatorname{tr}\mathfrak{e} + r_2 \operatorname{tr}\mathfrak{d}\mathfrak{e} + s_2 \operatorname{tr}\mathfrak{d}^2\mathfrak{e}, \end{cases} \tag{3.4}$$

where the quantities p, q, r, s and also the a_3, a_4, a_5 are polynomials in $\operatorname{tr}\mathfrak{d}$, $\operatorname{tr}\mathfrak{d}^2$, $\operatorname{tr}\mathfrak{d}^3$ with coefficients which are functions of the density ρ.

Now from (3.3) we have

$$\operatorname{tr}\mathbf{f} = 3a_0 + a_1 \operatorname{tr}\mathfrak{d} + a_2 \operatorname{tr}\mathfrak{d}^2 + a_3 \operatorname{tr}\mathfrak{e} + 2a_4 \operatorname{tr}\mathfrak{d}\mathfrak{e} + 2a_5 \operatorname{tr}\mathfrak{d}^2\mathfrak{e}, \tag{3.5}$$

$$\operatorname{tr}\mathfrak{d}\mathbf{f} = a_0 \operatorname{tr}\sigma + a_1 \operatorname{tr}\mathfrak{d}^2 + a_2 \operatorname{tr}\mathfrak{d}^3 + a_3 \operatorname{tr}\mathfrak{d}\mathfrak{e} + 2a_4 \operatorname{tr}\mathfrak{d}^2\mathfrak{e} + 2a_5 \operatorname{tr}\mathfrak{d}^3\mathfrak{e}, \tag{3.6}$$

$$\operatorname{tr}\mathfrak{d}^2\mathbf{f} = a_0 \operatorname{tr}\mathfrak{d}^2 + a_1 \operatorname{tr}\mathfrak{d}^3 + a_2 \operatorname{tr}\mathfrak{d}^4 + a_3 \operatorname{tr}\mathfrak{d}^2\mathfrak{e} + 2a_4 \operatorname{tr}\mathfrak{d}^3\mathfrak{e} + 2a_5 \operatorname{tr}\mathfrak{d}^4\mathfrak{e}. \tag{3.7}$$

But the quantities $\operatorname{tr}\mathfrak{d}^4$, $\operatorname{tr}\mathfrak{d}^3\mathfrak{e}$, and $\operatorname{tr}\mathfrak{d}^4\mathfrak{e}$ which occur in (3.6) and (3.7) can be expressed by means of certain auxiliary formulae in terms of the quantities of the set (3.2). These formulae will be derived in the following section.

[7] R. S. Rivlin, *Further remarks on the stress-deformation relations for isotropic materials*, J. Rational Mech. Anal., 4, 1955, pp. 681–702.

4. Auxiliary formulae. From the Hamilton-Cayley theorem we have the following matrix relation

$$\sigma^3 - (\operatorname{tr}\sigma)\sigma^2 + M\sigma - NI = 0, \tag{4.1}$$

where,

$$M = \tfrac{1}{2}[(\operatorname{tr}\sigma)^2 - \operatorname{tr}\sigma^2]; \; N = \tfrac{1}{3}\operatorname{tr}\sigma^3 - \tfrac{1}{2}\operatorname{tr}\sigma\operatorname{tr}\sigma^2 + \tfrac{1}{6}(\operatorname{tr}\sigma)^3.$$

Multiplying (4.1) by the matrix σ and taking the trace of the resulting expression we obtain an equation which can be written in the form

$$\operatorname{tr}\sigma^4 = \tfrac{1}{3}\operatorname{tr}\sigma\operatorname{tr}\sigma^3 - (\operatorname{tr}\sigma)^2\operatorname{tr}\sigma^2 + \tfrac{1}{2}(\operatorname{tr}\sigma^2)^2 + \tfrac{1}{8}(\operatorname{tr}\sigma)^4. \tag{4.2}$$

Similarly if we multiply (4.1) by the matrices ε and $d\varepsilon$ and then take the trace of each of the matrix expressions so obtained, we find

$$\operatorname{tr}\sigma^3\varepsilon = \operatorname{tr}\sigma\operatorname{tr}\sigma^2\varepsilon - M\operatorname{tr} d\varepsilon + N\operatorname{tr}\varepsilon, \tag{4.3}$$

$$\operatorname{tr}\sigma^4\varepsilon = \operatorname{tr}\sigma\operatorname{tr}\sigma^3\varepsilon - M\operatorname{tr}\sigma^2\varepsilon + N\operatorname{tr} d\varepsilon, \tag{4.4}$$

in which the quantity $\operatorname{tr}\sigma^3\varepsilon$ in the right member of the second equation can be eliminated by the first of these equations.

Let us now eliminate the quantities $\operatorname{tr}\sigma^4$, $\operatorname{tr}\sigma^3\varepsilon$, and $\operatorname{tr}\sigma^4\varepsilon$ from the right members of (3.6) and (3.7) by the substitutions (4.2), (4.3), and (4.4). Let us then substitute the above expressions for $\operatorname{tr}\mathbf{f}$, $\operatorname{tr}\mathbf{df}$, and $\operatorname{tr}\sigma^5\mathbf{f}$ into the equation (2.4). Making use of (3.4) the equation thus obtained can be written in the following form

$$P\operatorname{tr}\varepsilon + Q\operatorname{tr} d\varepsilon + R\operatorname{tr}\sigma^2\varepsilon + S = 0, \tag{4.5}$$

in which the quantities P, Q, R, and S are linear expressions in the derivatives of E with respect to ρ, θ, ξ, and ζ, and it is seen moreover that the coefficients of these expressions depend on the density ρ and the stress tensor σ. The explicit equations for P, Q, R, and S are

$$P = -\rho\frac{\partial E}{\partial\rho} + (a_3 + 3q_0 + q_1\theta + q_2\xi)\frac{\partial E}{\partial\theta} + \alpha\frac{\partial E}{\partial\xi} + A\frac{\partial E}{\partial\zeta}, \tag{4.6}$$

$$Q = -\frac{1}{\rho} + (2a_4 + 3r_0 + r_1\theta + r_2\xi)\frac{\partial E}{\partial\theta} + \beta\frac{\partial E}{\partial\xi} + B\frac{\partial E}{\partial\zeta}, \tag{4.7}$$

$$R = (2a_5 + 3s_0 + s_1\theta + s_2\xi)\frac{\partial E}{\partial\theta} + \gamma\frac{\partial E}{\partial\xi} + C\frac{\partial E}{\partial\zeta}, \tag{4.8}$$

$$S = (3p_0 + p_1\theta + p_2\xi)\frac{\partial E}{\partial\theta} + \delta\frac{\partial E}{\partial\xi} + D\frac{\partial E}{\partial\zeta}, \tag{4.9}$$

with,

$$\alpha = (\tfrac{4}{3}\zeta - 2\theta\xi + \tfrac{2}{3}\theta^3)a_6 + 2(q_0\theta + q_1\xi + q_2\zeta), \tag{4.10}$$

$$\beta = 2(a_3 - a_6\theta^2 + a_6\xi + r_0\theta + r_1\xi + r_2\zeta), \tag{4.11}$$

$$\gamma = 2(2a_4 + 2a_5\theta + s_0\theta + s_1\xi + s_2\zeta), \tag{4.12}$$

$$\delta = 2(p_0\theta + p_1\xi + p_2\zeta), \tag{4.13}$$

and,

$$A = (2\zeta - 3\theta\xi + \theta^3)a_4 + (2\theta\zeta - 3\theta^2\xi + \theta^4)a_5 + 3(q_0\xi + q_1\zeta) + Gq_2\,. \tag{4.14}$$

$$B = 3(\xi - \theta^2)a_4 + 2(\zeta - \theta^3)a_5 + 3(r_0\xi + r_1\zeta) + Gr_2\,, \tag{4.15}$$

$$C = 3a_3 + 6a_4\theta + 3(\xi + \theta^2)a_5 + 3(s_0\xi + s_1\zeta) + Gs_2\,, \tag{4.16}$$

$$D = 3(p_0\xi + p_1\zeta) + Gp_2\,, \tag{4.17}$$

where we have used the abbreviation

$$G = \tfrac{1}{2}\theta^4 - 3\theta^2\xi + 4\theta\zeta + \tfrac{3}{2}\xi^2.$$

We recall that the quantities p, q, r, s and the a_3, a_4, a_5 in these equations are polynomials in θ, ξ, and ζ with coefficients which are constants or depend at most on the density ρ.

5. A lemma on independent scalars. Assign values to the components of the stress tensor σ such that its principal values σ_i are distinct. Put $X_i = \epsilon_{ii}$ (not summed) and consider the following three identities

$$X_1 + X_2 + X_3 = \operatorname{tr} \epsilon,$$

$$\sigma_1 X_1 + \sigma_2 X_2 + \sigma_3 X_3 = \operatorname{tr} \sigma\epsilon,$$

$$\sigma_1^2 X_1 + \sigma_2^2 X_2 + \sigma_3^2 X_3 = \operatorname{tr} \sigma^2\epsilon,$$

in which X's are components of the tensor ϵ, as indicated, relative to the origin of a canonical coordinate system for the tensor σ i.e., a rectangular system whose axes coincide with the principal directions of this tensor at an arbitrary point P. Regarding these as equations for the determination of the quantities X_i the determinant Δ of the coefficients in the equations is

$$\Delta = \begin{vmatrix} 1 & 1 & 1 \\ \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1^2 & \sigma_2^2 & \sigma_3^2 \end{vmatrix} = (\sigma_1 - \sigma_2)(\sigma_2 - \sigma_3)(\sigma_3 - \sigma_1).$$

Hence $\Delta \neq 0$ since the σ_i are distinct by assumption and the above equations therefore have a solution X_i for arbitrarily assigned values of the right members of these equations. *In other words, if the components $\sigma_{\alpha\beta}$ are chosen so that their principal values σ_i are distinct the components $\epsilon_{\alpha\beta}$ can be selected in such a way that the quantities* tr ϵ, *tr* $\sigma\epsilon$, *and* tr $\sigma^2\epsilon$ *assume arbitrarily assigned values at any point P.*

It follows from the above result and the fact that the relation (4.5) is satisfied identically (see §1) that this relation must decompose into the following four equations

$$P = Q = R = S = 0, \tag{5.1}$$

provided the principal values of the stress tensor are distinct. However, it is clear that the validity of the equations (5.1) is independent of the condition on the principal values since these equations must likewise be satisfied identically in the density ρ and the components of the stress tensor.

It has now been shown that a scalar $E(\rho, \sigma)$ will represent the deformation energy of a material medium having a stress-strain matrix $\mathbf{f}$ *of the type specified by the equation* (3.3) *if, and only if, it satisfies the set of differential equations* (5.1).

6. Preferred systems. Let us now assume that the quantities a_2, a_4, and a_5 in (3.3) vanish so that the stress-strain matrix $\mathbf{f}$ becomes

$$\mathbf{f} = a_0\mathbf{I} + a_1\boldsymbol{\sigma} + a_3\boldsymbol{\varepsilon}. \tag{6.1}$$

Under this condition the differential equations (5.1) reduce to

$$P = -\rho\frac{\partial E}{\partial \rho} + (a_3 + 3q_0 + q_1\theta)\frac{\partial E}{\partial \theta} + \alpha\frac{\partial E}{\partial \xi} + A\frac{\partial E}{\partial \zeta} = 0, \tag{6.2}$$

$$Q = -\frac{1}{\rho} + (3r_0 + r_1\theta)\frac{\partial E}{\partial \theta} + \beta\frac{\partial E}{\partial \xi} + B\frac{\partial E}{\partial \zeta} = 0, \tag{6.3}$$

$$R = (3s_0 + s_1\theta)\frac{\partial E}{\partial \theta} + \gamma\frac{\partial E}{\partial \xi} + C\frac{\partial E}{\partial \zeta} = 0, \tag{6.4}$$

$$S = (3p_0 + p_1\theta)\frac{\partial E}{\partial \theta} + \delta\frac{\partial E}{\partial \xi} + D\frac{\partial E}{\partial \zeta} = 0, \tag{6.5}$$

where,

$$\begin{cases} \alpha = 2(q_0\theta + q_1\xi); & \beta = 2(a_3 + r_0\theta + r_1\xi), \\ \gamma = 2(s_0\theta + s_1\xi); & \delta = 2(p_0\theta + p_1\xi), \end{cases} \tag{6.6}$$

and,

$$\begin{cases} A = 3(q_0\xi + q_1\zeta); & B = 3(r_0\xi + r_1\zeta), \\ C = 3(a_3 + s_0\xi + s_1\zeta); & D = 3(p_0\xi + p_1\zeta). \end{cases} \tag{6.7}$$

Let us moreover assume that the stress-strain matrix $\mathbf{f}$ is *quadratic* in the stress components $\sigma_{\alpha\beta}$. Then the matrix $\mathbf{f}$, given by (6.1), will be said to be *preferred* and the associated system of stress-strain relations (1.4) will be called a *preferred system* (see §1).

For a preferred matrix $\mathbf{f}$ the quantity $s_1 = 0$ and we shall have

$$\begin{cases} a_0 = p_0 + q_0 \operatorname{tr}\boldsymbol{\varepsilon} + r_0 \operatorname{tr}\mathbf{d}\boldsymbol{\varepsilon} + s_0 \operatorname{tr}\mathbf{d}^2\boldsymbol{\varepsilon}, \\ a_1 = p_1 + q_1 \operatorname{tr}\boldsymbol{\varepsilon} + r_1 \operatorname{tr}\mathbf{d}\boldsymbol{\varepsilon}, \end{cases} \tag{6.8}$$

where the coefficients p, q, r, and s may be considered to be given by equations of the form[8]

[8] Whether or not we introduce the factor ρ in these equations is immaterial since the double subscript quantities can themselves be functions of ρ. However, in order that we may later take these latter quantities to be constants independent of ρ it is convenient to write the equations in the form (6.9). A similar remark applies to the equation (6.10).

$$(6.9)\qquad \begin{cases} \rho p_0 = p_{00} + p_{01}\theta + p_{02}\theta^2 + p_{03}\xi, \\ \rho q_0 = q_{00} + q_{01}\theta + q_{02}\theta^2 + q_{03}\xi, \\ \rho r_0 = r_{00} + r_{01}\theta; \qquad \rho s_0 = s_{00}; \qquad \rho r_1 = r_{10}, \\ \rho p_1 = p_{10} + p_{11}\theta; \qquad \rho q_1 = q_{10} + q_{11}\theta. \end{cases}$$

The quantities with double subscripts in these equations must either be constants or depend on the density ρ. Combining (6.8) and (6.9) we obtain the complete form of the expressions for the coefficients a_0 and a_1 in (6.1). For the determination of the coefficient a_3 of the preferred matrix $\mathbf{f}$ we have correspondingly

$$(6.10)\qquad \rho a_3 = \alpha_0 + \alpha_1\theta + \alpha_2\theta^2 + \alpha_3\xi,$$

in which the α's are at most functions of ρ.

7. Some special solutions. Consider the possibility of a solution E of the system (6.2), $\cdots$, (6.5) of the form

$$(7.1)\qquad E = \phi(\xi + k\theta^2),$$

where k is a constant. Since such a function E is independent of ρ and ζ the terms involving the derivatives $\partial E/\partial\rho$ and $\partial E/\partial\zeta$ can be omitted from the above system (6.2), $\cdots$, (6.5). Hence these equations become

$$(7.2)\qquad \begin{cases} [\alpha_0 + 3q_{00} + (\alpha_1 + 3q_{01} + q_{10})\theta + (\alpha_2 + 3q_{02} + q_{11})\theta^2 + (\alpha_3 + 3q_{03})\xi]\dfrac{\partial E}{\partial\theta} \\ \qquad + 2[q_{00}\theta + q_{01}\theta^2 + q_{02}\theta^3 + q_{10}\xi + (q_{03} + q_{11})\theta\xi]\dfrac{\partial E}{\partial\xi} = 0, \end{cases}$$

$$(7.3)\qquad \begin{cases} [3r_{00} + (3r_{01} + r_{10})\theta]\dfrac{\partial E}{\partial\theta} \\ \qquad + 2[\alpha_0 + (\alpha_1 + r_{00})\theta + (\alpha_2 + r_{01})\theta^2 + (\alpha_3 + r_{10})\xi]\dfrac{\partial E}{\partial\xi} = 1, \end{cases}$$

$$(7.4)\qquad \begin{cases} [3p_{00} + (3p_{01} + p_{10})\theta + (3p_{02} + p_{11})\theta^2 + 3p_{03}\xi]\dfrac{\partial E}{\partial\theta} \\ \qquad + 2[p_{00}\theta + p_{01}\theta^2 + p_{02}\theta^3 + p_{10}\xi + (p_{03} + p_{11})\theta\xi]\dfrac{\partial E}{\partial\xi} = 0, \end{cases}$$

$$(7.5)\qquad s_{00}\left(3\frac{\partial E}{\partial\theta} + 2\theta\frac{\partial E}{\partial\xi}\right) = 0.$$

While the equations (7.2), $\cdots$, (7.5) must be satisfied whenever E depends on the quantities θ and ξ only we shall limit our attention to the case for which E has the form (7.1).

We observe first that if the quantity s_{00} is different from zero the general solution of (7.5) is given by (7.1) with $k = -\frac{1}{3}$. If s_{00} vanishes we remove this restriction on the constant k.

From (7.1) and the three equations (7.2), (7.3), and (7.4) we now obtain

$$(7.6)\quad \begin{aligned} &k[\alpha_0 + 3q_{00} + (\alpha_1 + 3q_{01} + q_{10})\theta + (\alpha_2 + 3q_{02} + q_{11})\theta^2 + (\alpha_3 + 3q_{03})\xi]\theta \\ &\qquad + q_{00}\theta + q_{01}\theta^2 + q_{02}\theta^3 + q_{10}\xi + (q_{03} + q_{11})\theta\xi = 0, \end{aligned}$$

$$(7.7)\quad \begin{aligned} &\{k[3r_{00} + (3r_{01} + r_{10})\theta]\theta + [\alpha_0 + (\alpha_1 + r_{00})\theta \\ &\qquad + (\alpha_2 + r_{01})\theta^2 + (\alpha_3 + r_{10})\xi]\}\phi' = \tfrac{1}{2}, \end{aligned}$$

$$(7.8)\quad \begin{aligned} &k[3p_{00} + (3p_{01} + p_{10})\theta + (3p_{02} + p_{11})\theta^2 + 3p_{03}\xi]\theta \\ &\qquad + p_{00}\theta + p_{01}\theta^2 + p_{02}\theta^3 + p_{10}\xi + (p_{03} + p_{11})\theta\xi = 0, \end{aligned}$$

where ϕ' denotes the derivative of ϕ with respect to its argument. Equations (7.6) and (7.8) do not involve ϕ' and must moreover be satisfied identically in the quantities θ and ξ which can be regarded as independent variables. Hence, equating to zero the coefficients of different powers of θ and ξ in (7.6) and (7.8), we find

$$(7.9)\quad \begin{aligned} &(3k+1)q_{00} + k\alpha_0 = 0; \quad (3k+1)q_{01} + k(\alpha_1 + q_{10}) = 0, \\ &(3k+1)q_{02} + k(\alpha_2 + q_{11}) = 0; \quad (3k+1)q_{03} + k\alpha_3 + q_{11} = 0, \\ &(3k+1)p_{01} + kp_{10} = 0; \quad (3k+1)p_{02} + kp_{11} = 0, \\ &(3k+1)p_{03} + p_{11} = 0; \quad (3k+1)p_{00} = p_{10} = q_{10} = 0. \end{aligned}$$

Let us now put

$$Z = \xi + k\theta^2,$$

and let us then use this relation to eliminate the quantity ξ from equation (7.7) to obtain

$$\begin{aligned} &\{\alpha_0 + [(3k+1)r_{00} + \alpha_1]\theta + [(3k+1)r_{01} + \alpha_2 - k\alpha_3]\theta^2 \\ &\qquad + (\alpha_3 + r_{10})Z\}\phi'(Z) = \tfrac{1}{2}. \end{aligned}$$

But this equation must be satisfied identically in θ and Z, considered as independent variables, and hence must decompose into the following relations

$$(7.10)\qquad (3k+1)r_{00} + \alpha_1 = 0; \quad (3k+1)r_{01} + \alpha_2 - k\alpha_3 = 0,$$

$$(7.11)\qquad 2[\alpha_0 + (\alpha_3 + r_{10})Z]\phi'(Z) = 1.$$

It is evident from (7.11) that the quantities α_0 and $\alpha_3 + r_{10}$ cannot both be equal to zero. But if $k = -\frac{1}{3}$ we see from the first equation (7.9) that we must have $\alpha_0 = 0$. Hence the following four cases give all possibilities:

Case I	$k = -\frac{1}{3}$;	$\alpha_3 + r_{10} \neq 0$,	
Case II	$k \neq -\frac{1}{3}$;	$\alpha_0 \neq 0$;	$\alpha_3 + r_{10} \neq 0$,
Case III	$k \neq -\frac{1}{3}$;	$\alpha_0 \neq 0$;	$\alpha_3 + r_{10} = 0$,
Case IV	$k \neq -\frac{1}{3}$;	$\alpha_0 = 0$;	$\alpha_3 + r_{10} \neq 0$,

We shall treat each of these cases separately in the following sections.

8. Solution for Case I. From equations (7.9) and (7.10) we have

$$\alpha_0 = \alpha_1 = p_{10} = p_{11} = q_{10} = 0; \qquad \alpha_2 = -q_{11}; \qquad \alpha_3 = 3q_{11}. \tag{8.1}$$

No restriction is placed on the coefficient a_0 in (6.1) by these conditions. In other words a_0 is determined by the first equation (6.8) and the equations (6.9) in which the double subscript quantities are arbitrary constants or depend at most on the density ρ. However, we find that the coefficients a_1 and a_2 in (6.1) have the special forms

$$a_1 = \frac{q_{11}\theta}{\rho} \operatorname{tr} \varepsilon + \frac{r_{10}}{\rho} \operatorname{tr} \delta\varepsilon; \qquad a_2 = -\frac{q_{11}\theta^2}{\rho} + \frac{3q_{11}\xi}{\rho}.$$

Using the values of α_0 and α_2 in (8.1) and integrating the equation (7.11) it follows that the deformation energy E is given by

$$E = \frac{\log (\xi - \theta^2/3)}{2(3q_{11} + r_{10})} + \text{const.}$$

This formula shows that E becomes infinite for vanishing stress and hence Case I must be discarded on physical grounds.

9. Solution for Case II. This is the general case since none of the quantities on which case restrictions are imposed, are definitely specified. Let us put $q_{00} = \lambda$ and $\alpha_0 = 2\mu$. Then the first equation (7.9) is equivalent to any one of the following three relations

$$k = \frac{-\lambda}{3\lambda + 2\mu}; \qquad \frac{-k}{3k+1} = \frac{\lambda}{2\mu}; \qquad 3k + 1 = \frac{2\mu}{3\lambda + 2\mu}. \tag{9.1}$$

Using these relations we now find that the remaining conditions (7.9) and the conditions (7.10) can be written in the form

$$q_{01} = \frac{\lambda\alpha_1}{2\mu}; \qquad q_{02} = \frac{\lambda(\alpha_2 + q_{11})}{2\mu}; \qquad q_{03} = \frac{\lambda\alpha_3}{2\mu} - \frac{(3\lambda + 2\mu)q_{11}}{2\mu},$$

$$p_{02} = \frac{\lambda p_{11}}{2\mu}; \qquad p_{03} = \frac{-(3\lambda + 2\mu)p_{11}}{2\mu}; \qquad r_{00} = \frac{-(3\lambda + 2\mu)\alpha_1}{2\mu},$$

$$r_{01} = \frac{-\lambda\alpha_3}{2\mu} - \frac{(3\lambda + 2\mu)\alpha_2}{2\mu}; \qquad p_{00} = p_{01} = p_{10} = 0.$$

In addition to the above relations we must have $s_{00} = 0$ from the observation in §7. Hence from (6.9) we have

$$\begin{aligned}
\rho p_0 &= \frac{\lambda\beta_1}{2\mu}\theta^2 - \left(\frac{3\lambda + 2\mu}{2\mu}\right)\beta_1\xi; \\
\rho r_0 &= -\left(\frac{3\lambda + 2\mu}{2\mu}\right)\alpha_1 - \left[\frac{\lambda\alpha_2}{2\mu} + \left(\frac{3\lambda + 2\mu}{2\mu}\right)\alpha_2\right]\theta, \\
\rho q_0 &= \lambda + \frac{\lambda\alpha_1}{2\mu}\theta + \frac{\lambda}{2\mu}(\alpha_2 + \beta_2)\theta^2 + \left[\frac{\lambda\alpha_3}{2\mu} - \left(\frac{3\lambda + 2\mu}{2\mu}\right)\beta_2\right]\xi, \\
s_0 &= 0; \qquad \rho r_1 = \beta_2; \qquad \rho p_1 = \beta_1\theta; \qquad \rho q_1 = \beta_2\theta.
\end{aligned} \tag{9.2}$$

where we have substituted β_1, β_2, and β_3 for p_{11}, q_{11} and r_{10} respectively. Using these relations it follows from (6.1), (6.8), and (6.10) that the stress-strain matrix $\mathbf{f}$ is given by

$$\mathbf{f} = \{\lambda\beta_1\theta^2 - (3\lambda + 2\mu)\beta_1\xi + [2\mu\lambda + \lambda\alpha_1\theta + \lambda(\alpha_2 + \beta_2)\theta^2 + \lambda\alpha_3\xi$$

$$(9.3) \qquad - (3\lambda + 2\mu)\beta_2\xi]\ \mathrm{tr}\ \varepsilon - [(3\lambda + 2\mu)(\alpha_1 + \alpha_2\theta) + \lambda\alpha_3\theta]\ \mathrm{tr}\ d\varepsilon\}\ \frac{\mathbf{I}}{2\mu\rho}$$

$$+ (\beta_1\theta + \beta_2\theta\ \mathrm{tr}\ \varepsilon + \beta_3\ \mathrm{tr}\ d\varepsilon)\ \frac{\mathfrak{d}}{\rho} + (2\mu + \alpha_1\theta + \alpha_2\theta^2 + \alpha_3\xi)\ \frac{\varepsilon}{\rho}.$$

The quantities λ, μ and the α's and β's which appear in this expression for $\mathbf{f}$ are functions of ρ subject only to the conditions specifying this case and the requirement that the associated deformation energy E, obtained by integration of (7.11), is a function of $\xi + k\theta^2$ alone, *i.e.* independent of the density ρ.

Integrating the equation (7.11) we find that *the deformation energy E is given by*

$$(9.4) \qquad E = \frac{\log\left[1 + \dfrac{\alpha_3 + \beta_3}{2\mu}\left(\xi - \dfrac{\lambda\theta^2}{3\lambda + 2\mu}\right)\right]}{2(\alpha_3 + \beta_3)},$$

where λ, μ and $\alpha_3 + \beta_3$ must be constants, i.e. independent of ρ, provided the constant of integration is chosen so that $E = 0$ when the stress tensor vanishes. The requirement that the quantities λ, μ and $\alpha_3 + \beta_3$ be constants results from the fact that E has been assumed to be independent of the density.

Expanding the right member of (9.4) we have

$$(9.5) \qquad E = \frac{1}{4\mu}\left(\xi - \frac{\lambda\theta^2}{3\lambda + 2\mu}\right) - \frac{(\alpha_3 + \beta_3)}{16\mu^2}\left(\xi - \frac{\lambda\theta^2}{3\lambda + 2\mu}\right)^2 + \cdots.$$

The first term in the right member of this equation can be interpreted as representing the deformation energy per unit mass for small displacements. We observe that this term is formally the same as the well known expression for the deformation energy per unit volume in the ordinary or classical elasticity theory. The fact that the expression, in our case, gives energy per unit mass rather than unit volume arises from the formal introduction of the factor ρ in the left members of the equations (6.9) and (6.10)

Remark. If we take $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and also $\beta_1 = \beta_2 = 0$ but $\beta_3 = m \neq 0$ the matrix $\mathbf{f}$ assumes the following simple form

$$\mathbf{f} = \frac{\lambda}{\rho}\ \mathrm{tr}\ \varepsilon\mathbf{I} + \frac{m}{\rho}\ (\mathrm{tr}\ d\varepsilon)\mathfrak{d} + \frac{2\mu}{\rho}\ \varepsilon.$$

The formula (9.4) for E now becomes

$$E = \frac{1}{2m}\log\left[1 + \frac{m}{2\mu}\left(\xi - \frac{\lambda\theta^2}{3\lambda + 2\mu}\right)\right].$$

No effect is had, as a result of this specialization of the matrix $\mathbf{f}$, on the approximate expression for the deformation energy for small displacements which is given by the first term in the right member of (9.5).

10. Solution for Case III. Putting $\beta_3 = -\alpha_3$ in (9.2) these equations determine the quantities a_0 and a_1 by means of (6.8). Then using (6.8) and (6.10) we find the stress-strain matrix $\mathbf{f}$ in accordance with (6.1). The explicit expression for this matrix is given by

$$\begin{aligned}\mathbf{f} = \{&\lambda\beta_1\theta^2 - (3\lambda + 2\mu)\beta_1\xi + [2\mu\lambda + \lambda\alpha_1\theta + \lambda(\alpha_2 + \beta_2)\theta^2 + \lambda\alpha_3\xi \\ &- (3\lambda + 2\mu)\beta_2\xi]\ \mathrm{tr}\,\varepsilon - [(3\lambda + 2\mu)(\alpha_1 + \alpha_2\theta) + \lambda\alpha_3\theta]\ \mathrm{tr}\,\delta\varepsilon\}\ \frac{\mathbf{I}}{2\mu\rho} \\ &+ (\beta_1\theta + \beta_2\theta\ \mathrm{tr}\,\varepsilon - \alpha_3\ \mathrm{tr}\,\delta\varepsilon)\ \frac{\delta}{\rho} + (2\mu + \alpha_1\theta + \alpha_2\theta^2 + \alpha_3\xi)\ \frac{\varepsilon}{\rho}.\end{aligned} \tag{10.1}$$

For this case the differential equation (7.11) becomes

$$4\mu\,\frac{d\phi}{dZ} = 1.$$

Hence, by integration, we find

$$E = \frac{1}{4\mu}\left(\xi - \frac{\lambda\theta^2}{3\lambda + 2\mu}\right), \tag{10.2}$$

when we use the value of k given in §9 and choose the constant of integration so that E vanishes for zero stress. *Hence the exact expression* (10.2) *for E is formally the same as the expression for the deformation energy in the ordinary elasticity theory* (*see* §9).

Remark. If we put the three quantities α and the three quantities β equal to zero in (10.1) the expression for $\mathbf{f}$ takes the very simple form

$$\mathbf{f} = \frac{\lambda}{\rho}\mathbf{I} + \frac{2\mu}{\rho}\varepsilon.$$

However, this expression for $\mathbf{f}$ is not sufficiently general to describe the behavior of a material medium subject to large stresses, *e.g.* stresses which are sufficiently large as to induce a state of yield or plastic flow.

11. Solution for Case IV. To determine the stress-strain matrix $\mathbf{f}$ for this case we must go back to the relations (7.9) and (7.10) with $\alpha_0 = 0$ and we must also take $s_{00} = 0$ as observed in §7. From the conditions on the double subscript quantities resulting from (7.9) and (7.10) when $k \neq -\frac{1}{3}$ we can proceed to the determination of the quantities a_0, a_1 and a_3 and hence obtain the general expression for the matrix $\mathbf{f}$ by means of (6.1). For this case the differential equation (7.11) becomes

$$2(\alpha_3 + \beta_3)Z\,\frac{d\phi}{dZ} = 1,$$

where we have substituted β_3 for r_{10}. Integrating we therefore have

$$E = \frac{\log\,(\xi + k\theta^2)}{2(\alpha_3 + \beta_3)} + \text{const.},$$

where the quantity $\alpha_3 + \beta_3$ must be a constant, *i.e.* independent of the density ρ.

This deformation energy E does not remain finite for vanishing stress and hence this case fails to meet the requirements of the physical problem.

12. Deformation energies which depend on the density. It will now be convenient to replace the equations (6.9) by

$$\begin{cases} p_0 = p_{00} + p_{01}\theta + p_{02}\theta^2 + p_{03}\xi, \\ q_0 = q_{00} + q_{01}\theta + q_{02}\theta^2 + q_{03}\xi, \\ r_0 = r_{00} + r_{01}\theta; \qquad s_0 = s_{00}\,; \qquad r_1 = r_{10}\,, \\ p_1 = p_{10} + p_{11}\theta; \qquad q_1 = q_{10} + q_{11}\theta, \end{cases} \tag{12.1}$$

where the double subscript quantities are again either constants or at most functions of ρ. Similarly we replace (6.10) by

$$a_3 = \alpha_0 + \alpha_1\theta + \alpha_2\theta^2 + \alpha_3\xi, \tag{12.2}$$

with the α's constants or functions of ρ. Using (6.8), (12.1), and (12.2) we can now determine the general form of the preferred matrix $\mathbf{f}$ in accordance with (6.1).

Let us now consider the possibility of a solution E of the equations (6.2), $\cdots$, (6.5) of the form

$$E = \frac{\psi(\theta, \xi, \zeta)}{\rho}. \tag{12.3}$$

Making this substitution in the equations (6.2), $\cdots$, (6.5) we obtain

$$\begin{cases} (a_3 + 3q_0 + q_1\theta)\dfrac{\partial\psi}{\partial\theta} + \alpha\dfrac{\partial\psi}{\partial\xi} + A\dfrac{\partial\psi}{\partial\zeta} + \psi = 0, \\ (3r_0 + r_1\theta)\dfrac{\partial\psi}{\partial\theta} + \beta\dfrac{\partial\psi}{\partial\xi} + B\dfrac{\partial\psi}{\partial\zeta} = 1, \\ 3s_0\dfrac{\partial\psi}{\partial\theta} + \gamma\dfrac{\partial\psi}{\partial\xi} + C\dfrac{\partial\psi}{\partial\zeta} = 0, \\ (3p_0 + p_1\theta)\dfrac{\partial\psi}{\partial\theta} + \delta\dfrac{\partial\psi}{\partial\xi} + D\dfrac{\partial\psi}{\partial\zeta} = 0, \end{cases} \tag{12.4}$$

where the coefficients $\alpha, \beta, \gamma, \delta$ and A, B, C, D are given by (6.6) and (6.7) and where $s_1 = 0$ for the preferred system under consideration (see §6). We shall investigate a particular class of solutions ψ of these equations in the following section.

13. Other special solutions. Corresponding to the solutions considered in §7 let us now assume that ψ has the form

$$\psi = \psi(Z), \qquad Z = \xi + k\theta^2 \tag{13.1}$$

where k is a constant. Making this substitution the equations (12.4) immediately become

$$(13.2)\qquad \begin{cases} 2[k(a_2+3q_0+q_1\theta)\theta+(q_0\theta+q_1\xi)]\psi'+\psi = 0, \\ 2[k(3r_0+r_1\theta)\theta+(a_2+r_0\theta+r_1\xi)]\psi' = 1, \\ s_0(3k+1)\theta = 0; \quad k(3p_0+p_1\theta)\theta+p_0\theta+p_1\xi = 0, \end{cases}$$

where ψ' denotes the derivative of ψ with respect to its argument Z.

Now make the substitutions (12.1), and (12.2) in (13.2) and follow this by elimination of the quantity ξ from the resulting equations by means of the second relation (13.1). From the first equation (13.2) we thus obtain

$$(13.3)\qquad \begin{cases} \{k\alpha_0+(3k+1)q_{00}+[k\alpha_2+(3k+1)q_{02}+q_{11}]Z\}\theta\psi' \\ +[k\alpha_1+(3k+1)q_{01}]\theta^2\psi'+q_{10}Z\psi'+\tfrac{1}{2}\psi \\ +[k(\alpha_2-k\alpha_3)+(3k+1)(q_{02}-kq_{03})]\theta^3\psi' = 0. \end{cases}$$

But this equation must be satisfied identically in the independent variables θ and Z. Hence (13.3) must decompose into the differential relation

$$(13.4)\qquad 2q_{10}Z\psi'+\psi = 0,$$

and the following algebraic relations

$$(13.5)\qquad \begin{cases} k\alpha_0+(3k+1)q_{00} = 0; \quad k\alpha_1+(3k+1)q_{01} = 0, \\ k\alpha_3+(3k+1)q_{03}+q_{11} = 0, \\ k(\alpha_2-k\alpha_3)+(3k+1)(q_{02}-kq_{03}) = 0. \end{cases}$$

Similarly from the second equation (13.2) we obtain

$$\{[\alpha_1+(3k+1)r_{00}]\theta+[\alpha_2-k\alpha_3+(3k+1)r_{01}]\theta^2+[\alpha_0+(\alpha_3+r_{10})Z]\}\psi' = \tfrac{1}{2}.$$

This equation breaks up into the two relations

$$(13.6)\qquad \alpha_1+(3k+1)r_{00} = 0; \quad (\alpha_2-k\alpha_3)+(3k+1)r_{01} = 0,$$

and the differential equation

$$(13.7)\qquad 2[\alpha_0+(\alpha_3+r_{10})Z]\psi' = 1.$$

Finally from the remaining two relations (13.2) we obtain the conditions

$$(13.8)\qquad \begin{cases} (3k+1)s_{00} = 0; \quad (3k+1)p_{00} = 0; \quad (3k+1)p_{01}+kp_{10} = 0, \\ (3k+1)p_{02}+kp_{11} = 0; \quad (3k+1)p_{03}+p_{11} = 0. \end{cases}$$

The function ψ can readily be found from the two differential equations (13.4) and (13.7). In fact if we eliminate ψ' from these equations we obtain

$$(13.9)\qquad \psi = \frac{-q_{10}Z}{\alpha_0+(\alpha_3+r_{10})Z}.$$

Substituting this expression for ψ into either of the equations (13.4) or (13.7) we are led to the relation

$$(13.10)\qquad (1+2q_{10})\alpha_0+(\alpha_3+r_{10})Z = 0,$$

which must be satisfied identically in Z. Hence we must have

$$q_{10} = -\tfrac{1}{2}; \qquad \alpha_0 + r_{10} = 0. \tag{13.11}$$

The conditions determining the preferred matrix $\mathbf{f}$ are given by the equations (13.5), (13.6), and (13.8). To express these conditions in convenient form let us put, as before, $\alpha_0 = 2\mu$ and $q_{00} = \lambda$. Assuming $3k + 1 \neq 0$ the first equation (13.5) is equivalent to the determination of k by (9.1). Setting

$$\beta_0 = p_{10}; \qquad \beta_1 = p_{11}; \qquad \beta_2 = q_{11}; \qquad \beta_3 = r_{10}, \tag{13.12}$$

the remaining conditions can be expressed by writing

$$\begin{cases} q_{01} = \dfrac{\lambda\alpha_1}{2\mu}; \qquad q_{02} = \dfrac{\lambda}{2\mu}(\alpha_2 + \beta_2); \qquad q_{03} = \dfrac{\lambda\alpha_3}{2\mu} - \dfrac{(3\lambda + 2\mu)}{2\mu}\beta_2, \\ r_{00} = -\dfrac{(3\lambda + 2\mu)}{2\mu}\alpha_1; \quad r_{01} = -\dfrac{(3\lambda + 2\mu)}{2\mu}\alpha_2 - \dfrac{\lambda\alpha_2}{2\mu}; \quad p_{00} = s_{00} = 0, \\ p_{01} = \dfrac{\lambda\beta_0}{2\mu}; \qquad p_{02} = \dfrac{\lambda\beta_1}{2\mu}; \qquad p_{03} = \dfrac{(3\lambda + 2\mu)}{2\mu}\beta_1. \end{cases} \tag{13.13}$$

To determine the explicit form of the stress-strain matrix $\mathbf{f}$ we first find the quantities in the left members of (12.1). Using (13.11) and (13.13) we thus obtain

$$\begin{aligned} p_0 &= \frac{\lambda\beta_0}{2\mu}\theta + \frac{\lambda\beta_1}{2\mu}\theta^2 + \frac{(3\lambda + 2\mu)\beta_1}{2\mu}\xi, \\ q_0 &= \lambda + \frac{\lambda\alpha_1}{2\mu}\theta + \frac{\lambda}{2\mu}(\alpha_2 + \beta_2)\theta^2 + \left[\frac{\lambda\alpha_3}{2\mu} - \frac{(3\lambda + 2\mu)\beta_2}{2\mu}\right]\xi, \\ r_0 &= -\frac{(3\lambda + 2\mu)\alpha_1}{2\mu} - \frac{(3\lambda + 2\mu)\alpha_2}{2\mu}\theta - \frac{\lambda\alpha_3\theta}{2\mu}; \qquad s_0 = 0, \\ p_1 &= \beta_0 + \beta_1\theta; \qquad q_1 = -\frac{1}{2} + \beta_2\theta; \qquad r_1 = \beta_3 = -\alpha_3. \end{aligned}$$

Hence from (6.1), (6.8), and (12.2) the matrix $\mathbf{f}$ is given by

$$\begin{aligned} \mathbf{f} = {} & \{\lambda\beta_0\theta + \lambda\beta_1\theta^2 + (3\lambda + 2\mu)\beta_1\xi + [2\mu\lambda + \lambda\alpha_1\theta \\ & + \lambda(\alpha_2 + \beta_2)\theta^2 + \lambda\alpha_3\xi - (3\lambda + 2\mu)\beta_2\xi]\operatorname{tr}\varepsilon \\ & - [(3\lambda + 2\mu)\alpha_1 + (3\lambda + 2\mu)\alpha_2\theta + \lambda\alpha_3\theta]\operatorname{tr}\delta\varepsilon\}\frac{\mathbf{I}}{2\mu} \\ & + [\beta_0 + \beta_1\theta - (\tfrac{1}{2} - \beta_2\theta)\operatorname{tr}\varepsilon - \alpha_3\operatorname{tr}\delta\varepsilon]\,\delta \\ & + (2\mu + \alpha_1\theta + \alpha_2\theta^2 + \alpha_3\xi)\varepsilon. \end{aligned} \tag{13.14}$$

From (12.3), (13.9), and (13.11) we have

$$E = \frac{1}{4\mu\rho}\left(\xi - \frac{\lambda\theta^2}{3\lambda + 2\mu}\right). \tag{13.15}$$

Hence we can state the following result. *The stress-strain matrix* $\mathbf{f}$ *having the form* (13.14) *in which the* λ, μ *are constants and the quantities* α *and* β *are functions of the density* ρ *admits the deformation energy* E *given by* (13.15). The necessity for taking λ and μ to be constants independent of ρ arises from the fact that E has been assumed to have the form (12.3) in this discussion.

The above expression (13.15) for the deformation energy E is exactly the same as the expression obtained for the energy per unit mass in the classical elasticity theory.

Remark. Taking all the α's and β's equal to zero it follows from (13.14) that the simple stress-strain matrix

$$\mathbf{f} = \lambda\,(\mathrm{tr}\ e)\mathbf{I} - \tfrac{1}{3}\,(\mathrm{tr}\ e)\sigma + 2\mu e,$$

where λ and μ are constants, is associated with the deformation energy E given by (13.15).

If $k = -\frac{1}{3}$ the conditions (13.5), (13.6) and (13.8) yield

$$\alpha_0 = \alpha_1 = p_{10} = p_{11} = 0; \qquad \alpha_2 = -q_{11}\,; \qquad \alpha_3 = 3q_{11}\,.$$

But q_{10} cannot vanish from the equation (13.14). Hence we can solve this equation for ψ'. Substituting the result of this solution into the equation (13.7) in which $\alpha_0 = 0$ we obtain a relation which is inconsistent with the assumption that ψ depends on the quantity Z. Hence $k = -\frac{1}{3}$ does not permit a solution of the problem.

Graduate Institute for Mathematics and Mechanics
Indiana University, Bloomington, Indiana

(Received March 29, 1956)

A DERIVATION OF THE EQUATIONS OF SHELL THEORY FOR GENERAL ORTHOGONAL COORDINATES*

By James K. Knowles and Eric Reissner

1. Introduction. The object of the present paper is a derivation of the differential equations of the linear theory of thin elastic shells for an arbitrary orthogonal system of coordinates on the middle surface of the shell. In this derivation we are limiting ourselves to what is often designated as Love's first approximation theory. We are not concerned with a derivation of our results from a general tensor-analytical formulation of shell theory. Rather, we wish to generalize an earlier, rather simple and direct derivation [1] in which the coordinate curves on the middle surface were assumed to be lines of curvature. This earlier derivation is reproduced, in somewhat greater detail, in C. T. Wang's Applied Elasticity [2].

Our earlier derivation made use of the fact that the lines of curvature on the middle surface, together with the normals to the surface, form an orthogonal system of coordinates in space. All other orthogonal systems on the middle surface, together with the normals to the surface, form non-orthogonal systems of coordinates in space. The essential point in the following derivation is the avoidance of the use of oblique space coordinates and of the notion of stress and strain referred to such oblique coordinates. Among derivations which make use of oblique coordinates we may refer to two recent publications by Cohen [3] and Weibel [4]. Certain of our final formulas may be compared with corresponding formulas in [3] which contain more terms than are obtained by us on the basis of the assumptions implicit in Love's first approximation theory.

We do not in the present paper consider the extension of our developments which are necessary to make them apply to higher-approximation theories, such as the Flügge-Byrne Theory [5] or the theory considered by one of the present writers [6, 7] and by Naghdi [8]. It will, however, be apparent that such extensions are possible.

2. Coordinate systems for shell equations. Let (ξ_1, ξ_2) be an arbitrary set of orthogonal coordinates on a surface called the middle surface. Let $\mathbf{r}(\xi_1, \xi_2)$ be the position vector to any point on the middle surface. With the notation

$$\alpha_m^2 = \frac{\partial \mathbf{r}}{\partial \xi_m} \cdot \frac{\partial \mathbf{r}}{\partial \xi_m} \qquad m = 1, 2 \quad (1)$$

we define unit tangent vectors $\mathbf{t}_m$ to the coordinate curves by

$$\mathbf{t}_m = \frac{1}{\alpha_m} \frac{\partial \mathbf{r}}{\partial \xi_m} \quad (2)$$

Let $\mathbf{n} = \mathbf{t}_1 \times \mathbf{t}_2$ be the unit normal vector to the middle surface. If we further define quantities $1/R_{mn}$ by the formulas

* A report on work supported by the office of Naval Research under Contract NR 064-418 with the Massachusetts Institute of Technology.

$$\frac{1}{R_{mn}} = \frac{1}{\alpha_m \alpha_n} \frac{\partial \mathbf{n}}{\partial \xi_m} \cdot \frac{\partial \mathbf{r}}{\partial \xi_n} = -\frac{1}{\alpha_m \alpha_n} \mathbf{n} \cdot \frac{\partial^2 \mathbf{r}}{\partial \xi_m \partial \xi_n} \tag{3}$$

we have the following differentiation formulas for the unit vectors, in which $m \neq n$,

$$\left.\begin{aligned} \frac{\partial \mathbf{t}_m}{\partial \xi_m} &= -\frac{1}{\alpha_n} \frac{\partial \alpha_m}{\partial \xi_n} \mathbf{t}_n - \frac{\alpha_m}{R_{mm}} \mathbf{n} \\ \frac{\partial \mathbf{t}_m}{\partial \xi_n} &= \frac{1}{\alpha_m} \frac{\partial \alpha_n}{\partial \xi_m} \mathbf{t}_n - \frac{\alpha_n}{R_{mn}} \mathbf{n} \\ \frac{\partial \mathbf{n}}{\partial \xi_m} &= \frac{\alpha_m}{R_{mm}} \mathbf{t}_m + \frac{\alpha_m}{R_{mn}} \mathbf{t}_n \end{aligned}\right\} \tag{4}$$

If we now extend the system (ξ_1, ξ_2) to a system (ξ_1, ξ_2, ζ) in space by introducing the perpendicular distance ζ from the middle surface, we have for the position vector $\mathbf{R}$ to a point in space

$$\mathbf{R}(\xi_1, \xi_2, \zeta) = \mathbf{r}(\xi_1, \xi_2) + \zeta \mathbf{n}(\xi_1, \xi_2) \tag{5}$$

From (5) we obtain unit vectors $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{n}$ to the coordinate curves. On the basis of (4), (2) we have for $\mathbf{T}_1$, $\mathbf{T}_2$

$$\mathbf{T}_m = \frac{\alpha_m}{A_m} \left[\left(1 + \frac{\zeta}{R_{mm}} \right) \mathbf{t}_m + \frac{\zeta}{R_{mn}} \mathbf{t}_n \right] \tag{6a}$$

where

$$A_m = \left| \frac{\partial \mathbf{R}}{\partial \xi_m} \right| = \alpha_m \left[\left(1 + \frac{\zeta}{R_{mm}} \right)^2 + \frac{\zeta^2}{R_{mn}^2} \right]^{\frac{1}{2}} \tag{6b}$$

The system (ξ_1, ξ_2, ζ) is orthogonal provided $\mathbf{T}_1 \cdot \mathbf{T}_2 = 0$. From

$$\mathbf{T}_1 \cdot \mathbf{T}_2 = \frac{\zeta}{R_{12}} \left[2 + \zeta \left(\frac{1}{R_{11}} + \frac{1}{R_{22}} \right) \right] \frac{\alpha_1 \alpha_2}{A_1 A_2}$$

it follows that a necessary and sufficient condition for orthogonality is

$$\frac{1}{R_{12}} = 0$$

Let us assume that the system (ξ_1, ξ_2) is such that $1/R_{12} \neq 0$. We may then determine a system (ξ_1, ξ_{II}) for which $1/R_{1\,II} = 0$. To do this we write

$$\xi_m = \xi_m(\xi_1, \xi_{II})$$

and we let ϕ be the angle enclosed by the ξ_m-curves and the ξ_M-curves. We have four relations of the form

$$\alpha_1 \alpha_1 \cos \phi = \frac{\partial \mathbf{r}}{\partial \xi_1} \cdot \frac{\partial \mathbf{r}}{\partial \xi_1} = \left(\frac{\partial \mathbf{r}}{\partial \xi_1} \frac{\partial \xi_1}{\partial \xi_1} + \frac{\partial \mathbf{r}}{\partial \xi_2} \frac{\partial \xi_2}{\partial \xi_1} \right) \cdot \frac{\partial \mathbf{r}}{\partial \xi_1} = \alpha_1^2 \frac{\partial \xi_1}{\partial \xi_1}, \text{ etc.}$$

from which we obtain

$$\frac{\partial \xi_1}{\partial \xi_I} = \frac{\alpha_I}{\alpha_1} \cos \phi, \qquad \frac{\partial \xi_1}{\partial \xi_{II}} = -\frac{\alpha_{II}}{\alpha_1} \sin \phi,$$

$$\frac{\partial \xi_2}{\partial \xi_I} = \frac{\alpha_I}{\alpha_2} \sin \phi, \qquad \frac{\partial \xi_2}{\partial \xi_{II}} = \frac{\alpha_{II}}{\alpha_2} \cos \phi,$$

and therewith the differentiation formulas

$$\left.\begin{aligned} \frac{1}{\alpha_I}\frac{\partial}{\partial \xi_I} &= \frac{\cos\phi}{\alpha_1}\frac{\partial}{\partial \xi_1} + \frac{\sin\phi}{\alpha_2}\frac{\partial}{\partial \xi_2}, \\ \frac{1}{\alpha_{II}}\frac{\partial}{\partial \xi_{II}} &= \frac{-\sin\phi}{\alpha_1}\frac{\partial}{\partial \xi_1} + \frac{\cos\phi}{\alpha_2}\frac{\partial}{\partial \xi_2} \end{aligned}\right\} \tag{7}$$

We now consider the defining equations for the quantities $1/R$ in the (ξ_I, ξ_{II}) system,

$$\frac{1}{R_{MN}} = \frac{1}{\alpha_M \alpha_N} \frac{\partial \mathbf{n}}{\partial \xi_M} \cdot \frac{\partial \mathbf{r}}{\partial \xi_N}$$

Introduction of the differentiation formulas (7) into these equations gives as relations between the quantities $1/R_{MN}$ and $1/R_{mn}$

$$\left.\begin{aligned} \frac{1}{R_{I\,I}} &= \frac{\cos^2\phi}{R_{11}} + \frac{\sin^2\phi}{R_{22}} + \frac{2\sin\phi\cos\phi}{R_{12}} \\ \frac{1}{R_{II\,II}} &= \frac{\sin^2\phi}{R_{11}} + \frac{\cos^2\phi}{R_{22}} - \frac{2\sin\phi\cos\phi}{R_{12}} \\ \frac{1}{R_{1\,II}} &= \sin\phi\cos\phi\left(\frac{1}{R_{22}} - \frac{1}{R_{11}}\right) + \frac{\cos^2\phi - \sin^2\phi}{R_{12}} \end{aligned}\right\} \tag{8}$$

The system (ξ_I, ξ_{II}) will be such that $1/R_{1\,II} = 0$ if ϕ is determined from the relation

$$\tan 2\phi = \frac{2/R_{12}}{1/R_{11} - 1/R_{22}} \tag{9}$$

3. Equilibrium equations of shell theory. We define stress resultant, stress couple and load intensity vectors $\mathbf{N}$, $\mathbf{M}$, $\mathbf{p}$ and $\mathbf{m}$ relevant to the (ξ_1, ξ_2) system as follows:

$$\left.\begin{aligned} \mathbf{N}_m &= N_{m1}\mathbf{t}_1 + N_{m2}\mathbf{t}_2 + Q_m\mathbf{n} \\ \mathbf{M}_m &= -M_{m1}\mathbf{t}_2 + M_{m2}\mathbf{t}_1 \\ \mathbf{p} &= p_1\mathbf{t}_1 + p_2\mathbf{t}_2 + q\mathbf{n} \\ \mathbf{m} &= -m_1\mathbf{t}_2 + m_2\mathbf{t}_1 \end{aligned}\right\} \tag{10}$$

The vector equilibrium equations

$$\left.\begin{aligned} &\frac{\partial \alpha_2 \mathbf{N}_1}{\partial \xi_1} + \frac{\partial \alpha_1 \mathbf{N}_2}{\partial \xi_2} + \alpha_1\alpha_2\mathbf{p} = 0 \\ &\frac{\partial \alpha_2 \mathbf{M}_1}{\partial \xi_1} + \frac{\partial \alpha_1 \mathbf{M}_2}{\partial \xi_2} + \alpha_1\alpha_2(\mathbf{N}_1 \times \mathbf{t}_1 + \mathbf{N}_2 \times \mathbf{t}_2 + \mathbf{m}) = 0 \end{aligned}\right\} \tag{11}$$

for an element of the shell determined by surfaces ξ_m = constant and $\xi_m + d\xi_m$ = constant reduce to the following scalar form, if use is made of (4) and (10):

$$\left.\begin{aligned}
&\frac{\partial \alpha_2 N_{11}}{\partial \xi_1} + \frac{\partial \alpha_1 N_{21}}{\partial \xi_2} + \frac{\partial \alpha_1}{\partial \xi_2} N_{12} - \frac{\partial \alpha_2}{\partial \xi_1} N_{22} + \alpha_1 \alpha_2 \left(\frac{Q_1}{R_{11}} + \frac{Q_2}{R_{12}} + p_1 \right) = 0 \\
&\frac{\partial \alpha_2 N_{12}}{\partial \xi_1} + \frac{\partial \alpha_1 N_{22}}{\partial \xi_2} + \frac{\partial \alpha_2}{\partial \xi_1} N_{21} - \frac{\partial \alpha_1}{\partial \xi_2} N_{11} + \alpha_1 \alpha_2 \left(\frac{Q_1}{R_{12}} + \frac{Q_2}{R_{22}} + p_2 \right) = 0 \\
&\frac{\partial \alpha_2 Q_1}{\partial \xi_1} + \frac{\partial \alpha_1 Q_2}{\partial \xi_2} - \alpha_1 \alpha_2 \left(\frac{N_{11}}{R_{11}} + \frac{N_{22}}{R_{22}} + \frac{N_{12}}{R_{12}} + \frac{N_{21}}{R_{12}} \right) + \alpha_1 \alpha_2 q = 0 \\
&\frac{\partial \alpha_2 M_{11}}{\partial \xi_1} + \frac{\partial \alpha_1 M_{21}}{\partial \xi_2} + \frac{\partial \alpha_1}{\partial \xi_2} M_{12} - \frac{\partial \alpha_2}{\partial \xi_1} M_{22} - \alpha_1 \alpha_2 Q_1 + \alpha_1 \alpha_2 m_1 = 0 \\
&\frac{\partial \alpha_2 M_{12}}{\partial \xi_1} + \frac{\partial \alpha_1 M_{22}}{\partial \xi_2} + \frac{\partial \alpha_2}{\partial \xi_1} M_{21} - \frac{\partial \alpha_1}{\partial \xi_2} M_{11} - \alpha_1 \alpha_2 Q_2 + \alpha_1 \alpha_2 m_2 = 0 \\
&\frac{M_{12}}{R_{11}} - \frac{M_{21}}{R_{22}} + N_{12} - N_{21} + \frac{M_{22} - M_{11}}{R_{12}} = 0
\end{aligned}\right\} \tag{12}$$

Equations (12), which may also be found in [3], generalize corresponding equations for the special case $1/R_{12} = 0$ in [1, 2]. As in the case $1/R_{12} = 0$ we have that the last of the six equations (12) expresses moment equilibrium about the normal to the middle surface and must reduce to an identity if appropriate expressions for resultants and couples in terms of stresses are introduced.

We note that while it is readily possible to express resultants and couples in terms of stresses for the orthogonal system (ξ_1, ξ_{II}, ζ), we would need definitions for stresses with respect to non-orthogonal systems to do the same for the system (ξ_1, ξ_2, ζ), and this we wish to avoid.

4. Analysis of strain. As in [1], but with somewhat different notation, we begin with expressions for components of strain with respect to a general three-dimensional orthogonal system of coordinates (η_1, η_2, η_3). With $\mathbf{R} = \mathbf{R}(\eta_1, \eta_2, \eta_3)$, $B_m^2 = \partial \mathbf{R}/\partial \eta_m \cdot \partial \mathbf{R}/\partial \eta_m$ and $\mathbf{U}$ for the displacement vector we may write for components of strain Γ_{mn}

$$\Gamma_{mn} = \frac{1}{B_m B_n} \left(\frac{\partial \mathbf{R}}{\partial \eta_m} \cdot \frac{\partial \mathbf{U}}{\partial \eta_n} + \frac{\partial \mathbf{R}}{\partial \eta_n} \cdot \frac{\partial \mathbf{U}}{\partial \eta_m} \right) \qquad m, n = 1, 2, 3 \tag{13}$$

For the application to shell theory we take $\eta_1 = \xi_I$, $\eta_2 = \xi_{II}$, $\eta_3 = \zeta$, $B_1 = \alpha_I(1 + \zeta/R_{II})$, $B_2 = \alpha_{II}(1 + \zeta/R_{II\,II})$, $B_3 = 1$, and we assume for the displacement vector,

$$\mathbf{U} = \mathbf{u}(\xi_I, \xi_{II}) + \zeta \mathbf{u}'(\xi_I, \xi_{II}) \tag{14}$$

From (13) and (14) we have

$$\Gamma_{MN} = \frac{\left(\frac{\partial \mathbf{r}}{\partial \xi_M} + \zeta \frac{\partial \mathbf{n}}{\partial \xi_M} \right) \cdot \left(\frac{\partial \mathbf{u}}{\partial \xi_N} + \zeta \frac{\partial \mathbf{u}'}{\partial \xi_N} \right) + \left(\frac{\partial \mathbf{r}}{\partial \xi_N} + \zeta \frac{\partial \mathbf{n}}{\partial \xi_N} \right) \cdot \left(\frac{\partial \mathbf{u}}{\partial \xi_M} + \zeta \frac{\partial \mathbf{u}'}{\partial \xi_M} \right)}{\alpha_M \alpha_N \left(1 + \frac{\zeta}{R_{MM}} \right) \left(1 + \frac{\zeta}{R_{NN}} \right)} \tag{15}$$

$$\Gamma_{M\zeta} = \frac{\mathbf{n}\cdot\left(\frac{\partial \mathbf{u}}{\partial \xi_M} + \zeta\frac{\partial \mathbf{u}'}{\partial \xi_M}\right) + \left(\frac{\partial \mathbf{r}}{\partial \xi_M} + \zeta\frac{\partial \mathbf{n}}{\partial \xi_M}\right)\cdot\mathbf{u}'}{\alpha_M\left(1 + \frac{\zeta}{R_{MM}}\right)} \tag{16}$$

$$\Gamma_{\zeta\zeta} = \mathbf{n}\cdot\mathbf{u}' \tag{17}$$

or, in view of the differentiation formulas (4) specialized to the $(\xi_{\mathrm{I}}, \xi_{\mathrm{II}})$ system,

$$\Gamma_{MN} = \frac{\frac{\partial \mathbf{r}}{\partial \xi_M}\cdot\left(\frac{\partial \mathbf{u}}{\partial \xi_N} + \zeta\frac{\partial \mathbf{u}'}{\partial \xi_N}\right)}{\alpha_M\alpha_N\left(1 + \frac{\zeta}{R_{NN}}\right)} + \frac{\frac{\partial \mathbf{r}}{\partial \xi_N}\cdot\left(\frac{\partial \mathbf{u}}{\partial \xi_M} + \zeta\frac{\partial \mathbf{u}'}{\partial \xi_M}\right)}{\alpha_M\alpha_N\left(1 + \frac{\zeta}{R_{MM}}\right)} \tag{15'}$$

$$\Gamma_{M\zeta} = \frac{1}{\alpha_M}\frac{\partial \mathbf{r}}{\partial \xi_M}\cdot\mathbf{u}' + \mathbf{n}\cdot\frac{\left(\frac{\partial \mathbf{u}}{\partial \xi_M} + \zeta\frac{\partial \mathbf{u}'}{\partial \xi_M}\right)}{\alpha_M\left(1 + \frac{\zeta}{R_{MM}}\right)} \tag{16'}$$

We limit ourselves to Love's first approximation by setting

$$1 + \frac{\zeta}{R_{MM}} \cong 1 \tag{18}$$

in Eqs. (15′), (16′). We may then write (15′) in the form

$$\Gamma_{MN} = \gamma_{MN} + \zeta\gamma'_{MN} \tag{19}$$

where

$$\gamma_{MN} = \frac{1}{\alpha_M\alpha_N}\left(\frac{\partial \mathbf{r}}{\partial \xi_M}\cdot\frac{\partial \mathbf{u}}{\partial \xi_N} + \frac{\partial \mathbf{r}}{\partial \xi_N}\cdot\frac{\partial \mathbf{u}}{\partial \xi_M}\right) \tag{20}$$

and

$$\gamma'_{MN} = \frac{1}{\alpha_M\alpha_N}\left(\frac{\partial \mathbf{r}}{\partial \xi_M}\cdot\frac{\partial \mathbf{u}'}{\partial \xi_N} + \frac{\partial \mathbf{r}}{\partial \xi_N}\cdot\frac{\partial \mathbf{u}'}{\partial \xi_M}\right) \tag{21}$$

are expressions for middle surface strains and bending strains, respectively.

We now utilize the first part of the Euler-Bernoulli hypothesis, according to which normals to the middle surface are deformed without extension into normals by setting

$$\Gamma_{\zeta\zeta} = \mathbf{n}\cdot\mathbf{u}' = 0 \tag{22}$$

This enables us to write

$$\mathbf{n}\cdot\frac{\partial \mathbf{u}'}{\partial \xi_M} = \frac{\partial}{\partial \xi_M}(\mathbf{n}\cdot\mathbf{u}') - \frac{\partial \mathbf{n}}{\partial \xi_M}\cdot\mathbf{u}' = -\frac{\partial \mathbf{n}}{\partial \xi_M}\cdot\mathbf{u}' = -\frac{1}{R_{MM}}\frac{\partial \mathbf{r}}{\partial \xi_M}\cdot\mathbf{u}'$$

and with this we obtain from (16′) and (18) the following statement of the second part of the Euler-Bernoulli hypothesis which relates the vector $\mathbf{u}'$ to the vector $\mathbf{u}$.

$$\Gamma_{M\zeta} = \gamma_{M\zeta} = \frac{1}{\alpha_M}\frac{\partial \mathbf{r}}{\partial \xi_M}\cdot\mathbf{u}' + \frac{\mathbf{n}}{\alpha_M}\cdot\frac{\partial \mathbf{u}}{\partial \xi_M} = 0 \tag{23}$$

To relate the components of $\mathbf{u}'$ to those of $\mathbf{u}$ we make use of (23) which according to (25) is equivalent to $\gamma_{1\zeta} = \gamma_{2\zeta} = 0$. Setting $\gamma_{m\zeta}$ of (26) equal to zero and using (27) and (4), we have

$$u'_m = -\frac{\mathbf{n}}{\alpha_m} \cdot \frac{\partial \mathbf{u}}{\partial \xi_m} = \frac{u_m}{R_{mm}} + \frac{u_n}{R_{mn}} - \frac{1}{\alpha_m}\frac{\partial w}{\partial \xi_m}, \qquad m = 1, 2; n = 2, 1 \quad (29)$$

We note that while the above expressions for γ_{mn} coincide with corresponding expressions in [3], the same is not true for the quantities γ'_{mn} which here contain fewer terms than in [3].

5. Stress-strain relations. For the orthogonal system $(\xi_{\text{I}}, \xi_{\text{II}}, \zeta)$ we have the conventional relations between stresses and strains for an isotropic material obeying Hooke's law. From these we obtain by integration across the thickness h relations between resultants N_{MN} and couples M_{MN} on the one hand and direct strains γ_{MN} and bending strains γ'_{MN} of the middle surface on the other hand. If E, G, ν denote Young's modulus, shear modulus and Poisson's ratio respectively, these relations are [1, 2]:

$$\left.\begin{aligned} N_{\text{I I}} &= \frac{Eh}{1-\nu^2}\frac{1}{2}(\gamma_{\text{I I}} + \nu\gamma_{\text{II II}}) \\ N_{\text{II II}} &= \frac{Eh}{1-\nu^2}\frac{1}{2}(\gamma_{\text{II II}} + \nu\gamma_{\text{I I}}) \\ N_{\text{II I}} &= N_{\text{I II}} = Gh\,\gamma_{\text{I II}} \end{aligned}\right\} \quad (30)$$

$$\left.\begin{aligned} M_{\text{I I}} &= \frac{Eh^3}{12(1-\nu^2)}\frac{1}{2}(\gamma'_{\text{I I}} + \nu\gamma'_{\text{II II}}) \\ M_{\text{II II}} &= \frac{Eh^3}{12(1-\nu^2)}\frac{1}{2}(\gamma'_{\text{II II}} + \nu\gamma'_{\text{I I}}) \\ M_{\text{II I}} &= M_{\text{I II}} = \frac{Gh^3}{12}\gamma'_{\text{I II}} \end{aligned}\right\} \quad (31)$$

Derivation of (30) and (31) is based on the evaluation of integrals of the form

$$(N, M) = \int_{-\frac{1}{2}h}^{\frac{1}{2}h} \sigma(1 + \zeta/R)(1, \zeta)d\zeta$$

for resultants and couples in terms of stresses and on again setting $1 + \zeta/R \cong 1$ [1, 2].

Our main problem is to obtain, without using the oblique system (ξ_1, ξ_2, ζ), corresponding relations for resultants and couples defined with respect to the system (ξ_1, ξ_2) in terms of γ_{mn} and γ'_{mn}. In order to accomplish this we establish transformation formulas for resultants and couples by consideration of the force and moment equilibrium of suitable infinitesimal curvilinear coordinate triangles on the middle surface. In this way we obtain the formulas

We now transform Eqs. (20), (21), (23) by means of the differentiation formulas (7). In this way we obtain the formulas

$$\left.\begin{aligned} \gamma_{\mathrm{I\,I}} &= \cos^2\phi\,\gamma_{11} + \sin^2\phi\,\gamma_{22} + 2\sin\phi\cos\phi\,\gamma_{12} \\ \gamma_{\mathrm{II\,II}} &= \sin^2\phi\,\gamma_{11} + \cos^2\phi\,\gamma_{22} - 2\sin\phi\cos\phi\,\gamma_{12} \\ \gamma_{\mathrm{I\,II}} &= \sin\phi\cos\phi(\gamma_{22} - \gamma_{11}) + (\cos^2\phi - \sin^2\phi)\gamma_{12} \end{aligned}\right\} \tag{24}$$

with an analogous set of formulas for $\gamma'_{\mathrm{I\,I}}$, $\gamma'_{\mathrm{II\,II}}$, $\gamma'_{\mathrm{I\,II}}$, and

$$\left.\begin{aligned} \gamma_{1\xi} &= \cos\phi\,\gamma_{1\xi} + \sin\phi\,\gamma_{2\xi} \\ \gamma_{\mathrm{II}\xi} &= -\sin\phi\,\gamma_{1\xi} + \cos\phi\,\gamma_{2\xi} \end{aligned}\right\} \tag{25}$$

In these equations the quantities γ_{mn}, γ'_{mn} and $\gamma_{m\xi}$ are defined as follows in terms of displacements and displacement derivatives.

$$\left.\begin{aligned} \gamma_{mn} &= \frac{1}{\alpha_m\alpha_n}\left(\frac{\partial \mathbf{r}}{\partial\xi_m}\cdot\frac{\partial \mathbf{u}}{\partial\xi_n} + \frac{\partial \mathbf{r}}{\partial\xi_n}\cdot\frac{\partial \mathbf{u}}{\partial\xi_m}\right) \\ \gamma'_{mn} &= \frac{1}{\alpha_m\alpha_n}\left(\frac{\partial \mathbf{r}}{\partial\xi_m}\cdot\frac{\partial \mathbf{u}'}{\partial\xi_n} + \frac{\partial \mathbf{r}}{\partial\xi_n}\cdot\frac{\partial \mathbf{u}'}{\partial\xi_m}\right) \\ \gamma_{m\xi} &= \frac{1}{\alpha_m}\left(\frac{\partial \mathbf{r}}{\partial\xi_m}\cdot\mathbf{u}' + \mathbf{n}\cdot\frac{\partial \mathbf{u}}{\partial\xi_m}\right) \end{aligned}\right\} \tag{26}$$

The final form for the expressions for γ_{mn}, γ'_{mn} and $\gamma_{m\xi}$ as given by equations (26) depends on the component representation employed for the vectors $\mathbf{u}$ and $\mathbf{u}'$. Often it will be convenient to write

$$\left.\begin{aligned} \mathbf{u} &= u_1\mathbf{t}_1 + u_2\mathbf{t}_2 + w\mathbf{n} \\ \mathbf{u}' &= u'_1\mathbf{t}_1 + u'_2\mathbf{t}_2 \end{aligned}\right\} \tag{27}$$

but other representations may sometimes be of advantage.

For the sake of completeness we list below the expressions which follow if (27) is introduced into (26) and use is made of the differentiation formulas (4). We also introduce the notation which was employed in [1].

$$\left.\begin{aligned} \varepsilon_1^0 &= \tfrac{1}{2}\gamma_{11} = \frac{1}{\alpha_1}\frac{\partial u_1}{\partial\xi_1} + \frac{u_2}{\alpha_1\alpha_2}\frac{\partial\alpha_1}{\partial\xi_2} + \frac{w}{R_{11}} \\ \varepsilon_2^0 &= \tfrac{1}{2}\gamma_{22} = \frac{1}{\alpha_2}\frac{\partial u_2}{\partial\xi_2} + \frac{u_1}{\alpha_1\alpha_2}\frac{\partial\alpha_2}{\partial\xi_1} + \frac{w}{R_{22}} \\ \gamma_{12}^0 &= \gamma_{12} = \frac{\alpha_2}{\alpha_1}\frac{\partial}{\partial\xi_1}\left(\frac{u_2}{\alpha_2}\right) + \frac{\alpha_1}{\alpha_2}\frac{\partial}{\partial\xi_2}\left(\frac{u_1}{\alpha_1}\right) + \frac{2w}{R_{12}} \\ \kappa_1 &= \tfrac{1}{2}\gamma'_{11} = \frac{1}{\alpha_1}\frac{\partial u'_1}{\partial\xi_1} + \frac{u'_2}{\alpha_1\alpha_2}\frac{\partial\alpha_1}{\partial\xi_2} \\ \kappa_2 &= \tfrac{1}{2}\gamma'_{22} = \frac{1}{\alpha_2}\frac{\partial u'_2}{\partial\xi_2} + \frac{u'_1}{\alpha_1\alpha_2}\frac{\partial\alpha_2}{\partial\xi_1} \\ \tau &= \gamma'_{12} = \frac{\alpha_2}{\alpha_1}\frac{\partial}{\partial\xi_1}\left(\frac{u'_2}{\alpha_2}\right) + \frac{\alpha_1}{\alpha_2}\frac{\partial}{\partial\xi_2}\left(\frac{u'_1}{\alpha_1}\right) \end{aligned}\right\} \tag{28}$$

ON THE UNSTEADY MOTION OF A VISCOUS FLUID PAST A SEMI-INFINITE FLAT PLATE*

By G. F. Carrier and R. C. DiPrima

1. Introduction. The subject of how a boundary layer responds to fluctuations in the external flow about a steady mean has been discussed recently by several authors [6, 7, 8, 9 10]. One reason for an interest in such problems arises in a study of the "Rijke tube" phenomenon [6, 7, 11]. There it appears to be important to know the phase lag in the heat transfer from a heated gauze situated in a stream with an unsteady velocity. Lighthill [7] investigated the oscillating flow problem using boundary layer theory and the Karman-Pohlhausen method. It is apparent that his results can not be valid near the leading edge of the object (say a heated ribbon). However most of the heat release takes place at the leading edge; hence it appears desirable to study the behaviour of the velocity fluctuations near the leading edge.

In this paper the viscous incompressible flow past a semi-infinite flat plate is investigated when the flow at infinity is $U_0 + U_1 \exp(i\omega t)$. The analysis is based on a system of equations derived by a modified Oseen linearization of the equations of motion [1, 2]. It is found that when $(\omega\nu)^{\frac{1}{2}} \gg U_0$ and U_1 the flow field consists of the sum of a mean flow (due to U_0) and an oscillatory motion (due to $U_1 \exp(i\omega t)$). The phase advance in the oscillatory motion is $\pi/8$ at the leading edge and increases monotonically to $\pi/4$ in a distance approximately given by $(\omega/2\nu)^{\frac{1}{2}}x = 2.5$. The result for $(\omega/2\nu)^{\frac{1}{2}}x \geqq 2.5$ is a shear wave type solution and corresponds to an infinite lamina performing harmonic oscillations in a viscous fluid.

In Sec. 5 the problem is considered when the interaction of the mean motion and the oscillatory motion can not be neglected. Expressions for the amplitude and phase advance of the unsteady skin friction near the leading edge are found. In particular the phase advance at the leading edge increases monotonically from 0 to $\pi/8$ as the dimensionless parameter $(\omega\nu)^{\frac{1}{2}}/(cU_0)$ increases from 0 to ∞. Here c is a constant which is defined in Sec. 2. Far down the plate the solution approaches the shear wave solution.

2. Formulation of Problem. The equations governing a two-dimensional viscous, incompressible flow are

$$\left.\begin{aligned} u_x + v_y &= 0, \\ \rho(u_t + uu_x + vu_y) &= -p_x + \mu\Delta u, \\ \rho(v_t + uv_x + vv_y) &= -p_y + \mu\Delta v. \end{aligned}\right\} \qquad (2.1)$$

* This work was sponsored by the Office of Naval Research under Contract No. 5-ORI-07666.

$$\left.\begin{aligned} N_{11} &= \cos^2\phi\, N_{\mathrm{I\,I}} + \sin^2\phi\, N_{\mathrm{II\,II}} - 2\sin\phi\cos\phi\, N_{\mathrm{I\,II}} \\ N_{22} &= \sin^2\phi\, N_{\mathrm{I\,I}} + \cos^2\phi\, N_{\mathrm{II\,II}} + 2\sin\phi\cos\phi\, N_{\mathrm{I\,II}} \\ N_{21} = N_{12} &= (N_{\mathrm{I\,I}} - N_{\mathrm{II\,II}})\sin\phi\cos\phi + N_{\mathrm{I\,II}}(\cos^2\phi - \sin^2\phi) \end{aligned}\right\} \tag{32}$$

and analogous relations between the M_{mn} and the M_{MN}.

We now introduce (30) into (32) to express the N_{mn} in terms of the γ_{MN}, and then use (24) to express the γ_{MN} in terms of the γ_{mn}. An analogous procedure leads to expressions for the M_{mn} in terms of the γ'_{mn}. The resulting equations represent the desired form of the stress-strain relations for an arbitrary orthogonal coordinate system on the middle surface of the shell:

$$\left.\begin{aligned} N_{11} &= \frac{Eh}{1-\nu^2}\frac{1}{2}(\gamma_{11} + \nu\gamma_{22}) \\ N_{22} &= \frac{Eh}{1-\nu^2}\frac{1}{2}(\gamma_{22} + \nu\gamma_{11}) \\ N_{21} = N_{12} &= Gh\,\gamma_{12} \end{aligned}\right\} \tag{33}$$

$$\left.\begin{aligned} M_{11} &= \frac{Eh^3}{12(1-\nu^2)}\frac{1}{2}(\gamma'_{11} + \nu\gamma'_{22}) \\ M_{22} &= \frac{Eh^3}{12(1-\nu^2)}\frac{1}{2}(\gamma'_{22} + \nu\gamma'_{11}) \\ M_{21} = M_{12} &= \frac{Gh^3}{12}\gamma'_{12} \end{aligned}\right\} \tag{34}$$

These relations have the same form as those for the special case of the lines of curvature coordinate system $(\xi_{\mathrm{I}}, \xi_{\mathrm{II}})$.

BIBLIOGRAPHY

1. Reissner, E. *A new derivation of the equations for the deformation of elastic shells.* Amer. J. Math. *63*, 177–184, 1941.
2. Wang, C. T. *Applied Elasticity*, New York, 1953, Chap. 12.
3. Cohen, J. W. *On stress calculations in helicoidal shells and propeller blades.* Report No. 21S, Netherlands Research Center T. N. O. for Shipbuilding and Navigation, Delft, Holland, July, 1955.
4. Weibel, E. S. *The strains and the energy in thin elastic shells of arbitrary shape for arbitrary deformation.* Z. für Ang. Math. u. Phys. *6*, 153–189, 1955.
5. Byrne, R. *Theory of small deformations of thin elastic shells*, University of California Publications in Mathematics, N. S. *2*, 103–152, 1944.
6. Reissner, E. *Note on the expressions for strains in bent thin shells.* Amer. J. Math. *64*, 768–772, 1942.
7. Reissner, E. *Stress-strain relations in the theory of thin shells.* J. Math. and Phys. *31*, 109–119, 1952.
8. Naghdi, P. M. *On the theory of thin elastic shells.* Q. Appl. Math. *14*, 369–380, 1956

Massachusetts Institute of Technology

(Received April 26, 1956)

Here x and y are the usual Cartesian coordinates, u and v are the corresponding velocities, p is the pressure, ρ the density (which is taken constant), $\nu = \mu/\rho$ is the kinematic viscosity, Δ is the Laplacian operator, and x, y, or t used as a subscript denotes partial differentiation with respect to that variable. We wish to determine the velocity field associated with a semi-infinite flat plate situated in an unsteady stream. For convenience we shall assume that the semi-infinite flat plate occupies the half line $y = 0$, $x \geqq 0$. In particular we wish to consider this problem when the imposed velocity, $U(t)$, far from the plate is of the form

$$U(t) = U_0 + U_1 e^{i\omega t}, \tag{2.2}$$

where U_0 and U_1 are constants. The boundary conditions which u and v must satisfy are

$$\begin{aligned} u = 0, \quad v = 0 \quad \text{on} \quad y = 0, \quad x \geqq 0, \\ u \to U_0 + U_1 e^{i\omega t}, \quad v \to 0 \end{aligned} \tag{2.3}$$

far from the plate.

In order to treat this problem we linearize Eqs. (2.1) by replacing the convective terms $uu_x + vu_y$, and $uu_x + vv_y$ by cUu_x and cUv_x. Here c is a constant which in general should depend upon the parameters involved in the problem. The case $U_1 \equiv 0$, $c = 1$ gives the well known Oseen approximation. The modification of the Oseen method (i.e. $c \neq 1$) has been used in treating the steady flow over a semi-infinite plate [1]. Also it has been applied to several other steady flow problems, and its rationalization discussed in [2]. In [2] it was noted that for small Reynolds numbers that the value $c = .43$ gave results in good agreement with experiment in all cases considered. When the flow is unsteady, the choice of c is clearly more complicated. However in a recent publication [3] this type of linearization was used to compute the fluctuating heat release from a heated ribbon when $U(t)$ had the form given by Eq. (2.2), and the choice of $c = .43$ gave excellent results.

The form of $U(t)$ suggests that we write u and v as

$$\left.\begin{aligned} u(x, y, t) &= u_0(x, y) + u_1(x, y)e^{i\omega t}, \\ &= U_0\{1 + \chi_y(x, y)\} + U_1\{1 + \psi_y(x, y)\}e^{i\omega t}, \end{aligned}\right\} \tag{2.4}$$

and

$$\left.\begin{aligned} v(x, y, t) &= v_0(x, y) + v_1(x, y)e^{i\omega t}, \\ &= -U_0\chi_x(x, y) - U_1\psi_x(x, y)e^{i\omega t}. \end{aligned}\right\}$$

In order to obtain a solution of this form it is necessary to neglect quadratic terms in u_1 ; i.e. we must restrict ourselves to cases in which U_1 is small compared to either U_0 or $(\omega\nu)^{\frac{1}{2}}$. Then using the linearization discussed above and Eqs. (2.4) we obtain the following equations for χ and ψ,

$$\Delta\Delta\chi - a\Delta\chi_x = 0, \tag{2.5}$$

$$\Delta\Delta\psi - a\Delta\psi_x - \beta^2\Delta\psi = a\Delta\chi_x. \tag{2.6}$$

Here $a = cU_0/\nu$, and $\beta^2 = i\omega/\nu$. The boundary conditions that χ and ψ must satisfy are

$$\left.\begin{aligned} &\chi_x = 0, \quad \psi_x = 0; \quad \chi_y = -1, \quad \psi_y = -1 \quad \text{on} \quad y = 0, \quad 0 \leqq x \leqq \infty, \\ &\chi_x \to 0, \quad \chi_y \to 0, \quad \psi_x \to 0, \quad \psi_y \to 0, \end{aligned}\right\} \tag{2.7}$$

far from the plate. The conditions ψ_x and $\chi_x = 0$ on $y = 0$, $0 \leqq x \leqq \infty$ clearly hold for all x and we satisfy them by requiring ψ and χ be constant on $y = 0$; in particular $\psi(x, 0) = \chi(x, 0) = 0$. As we shall see, when ω is very large the terms involving a in Eq. (2.6) may be neglected and we obtain homogeneous, independent equations for χ and ψ. That is to say, the velocity distribution is simply the flow due to a velocity U_0 at infinity plus that due to a velocity $U_1 e^{i\omega t}$ at infinity. In general of course these terms cannot be neglected, and hence the interaction of the mean flow and the time dependent flow must be considered. The term $a\Delta\psi_x$ represents the convective effect of the mean flow on the time dependent flow, while the non-homogeneous term $a\Delta\chi_x$ represents the convective effect of the time dependent flow on the mean flow.

To complete the formulation of our problem we must note that while χ, ψ, χ_y, ψ_y, χ_{yyy} and ψ_{yyy} are continuous across the plate χ_{yy} and ψ_{yy} which are proportional to the velocity gradient are discontinuous. We may express this discontinuity as

$$\chi_{yy}(x, 0+) = -\chi_{yy}(x, 0-) = \tfrac{1}{2}f_0(x),$$
$$\psi_{yy}(x, 0+) = -\psi_{yy}(x, 0-) = \tfrac{1}{2}f_1(x).$$

The functions $f_0(x)$ and $f_1(x)$ are identically zero for $x < 0$. Now the skin friction, τ, is given by

$$\tau = \mu \frac{\partial u}{\partial y}\bigg|_{y=0} = \mu \left\{\frac{\partial u_0}{\partial y} + \frac{\partial u_1}{\partial y} e^{i\omega t}\right\}_{y=0}.$$

Upon substituting for u_0 and u_1 from Eqs. (2.4) and using the above relations we may express τ in the form

$$\tau = \tau_0 + \tau_1 e^{i\omega t}, = \tfrac{1}{2}\mu U_0 f_0(x) + \tfrac{1}{2}\mu U_1 f_1(x) e^{i\omega t}. \tag{2.8}$$

In what is to follow we shall be primarily concerned with the determination of $f_0(x)$ and $f_1(x)$.

3. Case $U_1 = 0$. In this case our problem reduces to a solution of the problem of the flow over a semi-infinite flat plate due to a flow U_0 at infinity. Thus we must solve Eq. (2.5) subject to the boundary conditions (2.7). Even when $U_1 \neq 0$ we still need the solution of this problem in order to obtain ψ (see Eq. (2.6)). The solution of this problem has been given in [1]. Since the analysis of the general problem will follow the lines of [1], but is much more complicated in its detail; and also since we need the results found in [1] it is convenient to use this simpler problem to illustrate the method of analysis that is to be used. Hence we shall repeat briefly part of the computations described in [1].

Our problem can be most conveniently treated by the use of Fourier transforms. We define

$$\bar{\chi}(\xi, \eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(\xi x+\eta y)}\chi(x, y)\, dx\, dy, \qquad \bar{f}_0(\xi) = \int_{-\infty}^{\infty} e^{-i\xi x} f_0(x)\, dx. \tag{3.1}$$

In all that is to follow a barred quantity will denote the Fourier transform of that quantity. Taking the Fourier transform of Eq. (2.5), and using the boundary conditions and discontinuity relation for χ we obtain

$$(\xi^2 + \eta^2)(\xi^2 + \eta^2 + ia\xi)\bar{\chi}(\xi, \eta) = i\eta\bar{f}_0(\xi). \tag{3.2}$$

In order to avoid difficulties later it is desirable to write this equation as the limit as $k \to 0$ of the equation*

$$\{\xi^2 + \eta^2 + k^2\}\{\eta^2 + (\xi + ia)(\xi - ik)\}\bar{\chi}(\xi, \eta) = i\eta\bar{f}_0(\xi). \tag{3.3}$$

If we let $\bar{\chi}(\xi, y)$† denote the inverse transform of $\bar{\chi}(\xi, \eta)$ over η we obtain upon solving Eq. (3.3) and using simple contour integration that

$$\bar{\chi}(\xi, y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\eta y}\bar{\chi}(\xi, \eta)\, d\eta = \int_{-\infty}^{\infty} e^{-i\xi x}\chi(x, y)\, dx, \tag{3.4a}$$

$$= -\frac{y}{|y|}\,\frac{\exp\{-|y|(\zeta^2 + k^2)^{\frac{1}{2}}\} - \exp\{-|y|(\xi + ia)^{\frac{1}{2}}(\xi - ik)^{\frac{1}{2}}\}}{2i(a - k)(\zeta - ik)}\,\bar{f}_0(\xi) \tag{3.4b}$$

In particular we obtain from Eq. (3.4b) that

$$\bar{\chi}_y(\xi, y = 0) = \frac{-\bar{f}_0(\xi)/2}{(\xi - ik)^{\frac{1}{2}}\{(\xi + ia)^{\frac{1}{2}} + (\xi + ik)^{\frac{1}{2}}\}}\,. \tag{3.5}$$

A second equation for $\bar{\chi}_y(\xi, y = 0)$ can be obtained by using our boundary conditions. First however the condition that $\chi_y = -1$ on the plate must be written as the limit, as $\alpha \to 0$, of $\chi_y(x, 0) = -\exp(-\alpha x)$ on $x \geqq 0$. (The remarks made about k previously also apply to α.) Now differentiating (3.4a) with respect to y and setting $y = 0$ we have

$$\bar{\chi}_y(\xi, y = 0) = \int_{-\infty}^{\infty} e^{-i\xi x}\,\chi_y(x, y = 0)\, dx = \bar{v}(\xi) - \frac{1}{i(\xi - i\alpha)}\,, \tag{3.6}$$

where

$$\bar{v}(\xi) = \int_{-\infty}^{0} e^{-i\xi x}\chi_y(x, y = 0)\, dx.$$

The function $v(\xi)$ is of course not known. Equating (3.6) and (3.5) we obtain

* Here and throughout the rest of this paper the symbol k will be used to represent a small, positive, real quantity which is to be shrunk to zero at an appropriate moment.

† We shall be careful to write $\bar{\chi}(\xi, y)$ if we mean the x transform of $\chi(x, y)$ and $\bar{\chi}(\xi, \eta)$ if we are referring to the two-dimensional transform.

$$\bar{v}(\xi) - \frac{1}{i(\xi - i\alpha)} = \frac{-\bar{f}_0(\xi)/2}{(\xi + ik)^{\frac{1}{2}}\{(\xi + ik)^{\frac{1}{2}} + (\xi + i\alpha)^{\frac{1}{2}}\}} \tag{3.7}$$

We can now determine $\bar{v}(\xi)$ and $\bar{f}_0(\xi)/2$ by the Wiener-Hopf technique. If we anticipate that $\chi_y(x, y = 0)$ decays exponentially in $(-x)$ on the negative real axis, then $\bar{v}(\xi)$ is an analytic function of the complex variable ξ in some upper half-plane (UHP) which includes the real axis. Similarly $\bar{f}_0(\xi)$ is analytic in some lower half-plane (LHP) including the real axis. Also $1/(\xi - i\alpha)$ is analytic in the LHP, $\mathrm{Im}(\xi) < \alpha$. And finally $K(\xi) = (\xi - ik)^{\frac{1}{2}}\{(\xi + ik)^{\frac{1}{2}} + (\xi + i\alpha)^{\frac{1}{2}}\}$ is analytic in the strip $-k < \mathrm{Im}(\xi) < k$. Let us assume that we can split $K(\xi)$ into the product of two factors $K_+(\xi)$ and $K_-(\xi)$ such that $K_+(\xi)$ is analytic in the UHP, $\mathrm{Im}(\xi) > -k$; and $K_-(\xi)$ is analytic in the LHP $Im(\xi) < k$. Then we may rewrite Eq. (3.7) as

$$\frac{\bar{f}_0(\xi)/2}{K_-(\xi)} - \frac{K_+(i\alpha)}{i(\xi - i\alpha)} = -\bar{v}(\xi)K_+(\xi) - \frac{K_+(i\alpha) - K_+(\xi)}{i(\xi - i\alpha)} = E(\xi) \tag{3.8}$$

The left-hand of this equation is analytic in the LHP, while the right side is analytic in an overlapping UHP, and hence they define an entire function $E(\xi)$. By using order conditions of the known functions at infinity and the fact that $E(\xi)$ is an entire function we can show that $E(\xi) \equiv 0$, hence

$$\frac{\bar{f}_0(\xi)}{2} = \frac{K_+(i\alpha)K_-(\xi)}{i(\xi - i\alpha)} \tag{3.9}$$

We recall once again that we are interested in evaluating $\bar{f}_0(\xi)/2$ in the limit $\alpha \to 0$, $k \to 0$.

In this particular problem $K_-(\xi)$ and $K_+(0)$ can be determined by inspection of $K(\xi)$ (we will not be so fortunate later on). We see that $K_-(\xi) = (\xi - ik)^{\frac{1}{2}}$ and $K_+(0) = (i\alpha)^{\frac{1}{2}}$ in the limit as α and $k \to 0$. Substituting in (3.9) we may write $\bar{f}_0(\xi)$ as

$$\tfrac{1}{2}\bar{f}_0(\xi) = \frac{a^{\frac{1}{2}}}{(i\xi + k)^{\frac{1}{2}}} \tag{3.10}$$

We have written (3.10) in the preceding manner to indicate clearly that the singularity is in the UHP. In the limit as $k \to 0$ we find by the conventional inversion formula that

$$\tfrac{1}{2}f_0(x) = \begin{cases} \left(\dfrac{a}{\pi x}\right)^{\frac{1}{2}} = \left(\dfrac{cU_0}{\pi \nu x}\right)^{\frac{1}{2}} & x > 0, \\ 0 & x < 0. \end{cases} \tag{3.11}$$

As mentioned previously the choice of $c = .43$ gives good agreement with the non-linear theory and with experimental results for small Reynolds number. However in this case the Reynolds number is infinite and the result which is consistent with the Blasius solution is given by $c = .35$ [1]. It is clear that the numerical choice of c can not affect the character of the result, and hence for any value of c in this range the result will describe the basic features of the flow. Finally knowing $\bar{f}_0(\xi)$ it is possible to invert $\bar{\chi}(\xi, y)$ and obtain

$$\chi(x, y) = \frac{U_0}{\nu}\left\{\frac{2(r+x)^{\frac{1}{2}}}{(2\pi a)^{\frac{1}{2}}}\left[1 - e^{-a(x-r)/2}\right] + y\,\mathrm{erfc}\left[\frac{a(r-x)}{2}\right]^{\frac{1}{2}}\right\}^*,$$

where $r^2 = x^2 + y^2$.

From the preceding results and the definition of τ_0, Eq. (2.8), we have

$$\tau_0 = \mu U_0\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} \tag{3.12}$$

We can use Eq. (3.12) to determine the skin friction when $U_1 \neq 0$, by expanding τ_0 in a Taylor series about U_0. We have, if $\tau(U_0 + U_1)$ denotes the skin friction for a flow $U_0 + U_1$ at infinity that

$$\tau(U_0 + U_1) = \tau_0 + U_1\left.\frac{\partial \tau_0}{\partial U_0}\right|_{U_0} + \cdots = \left(1 + \frac{3}{2}\frac{U_1}{U_0}\right)\tau_0 + \cdots$$

This result is correct up to quadrate terms in U_1, and since this is the same as the accuracy in Eq. (2.6) we may use this result to compute τ_1 when $\omega = 0$. We have

$$\tau_1(\omega = 0) = \frac{3}{2}\mu U_1\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}}, \qquad \frac{1}{2}f_1(x, \omega = 0) = \frac{3}{2}\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} \tag{3.13}$$

These results will be of interest in the discussion in Sec. 5.

4. Case $U_0 = 0$. In this section we require that $U_1 \ll (\omega\nu)^{\frac{1}{2}}$ as indicated just after Eqs. (2.4). When $U_0 = 0$, Eq. (2.6) reduces to

$$\Delta\Delta\psi - \beta^2\Delta\psi = 0. \tag{4.1}$$

We should note that this equation gives the first approximation to the time dependent flow even when $U_0 \neq 0$ if the parameter $\sigma = (\omega\nu)^{\frac{1}{2}}/cU_0$ is large. This can be seen by introducing appropriate dimensionless variables in Eqs. (2.5) and (2.6) and then expanding ψ in powers of σ. The length scale that one should use is $(\nu/\omega)^{\frac{1}{2}}$.

To solve (4.1) we proceed in exactly the same manner as described in Sec. 3. The Fourier transform of Eq. (4.1) is

$$(\xi^2 + \eta^2 + \beta^2)(\xi^2 + \eta^2 + k^2)\bar{\psi}(\xi, \eta) = i\eta\bar{f}_{10}(\xi). \tag{4.2}$$

We have written $\bar{f}_{10}(\xi)$ instead of $\bar{f}_1(\xi)$ to indicate that this is the special case $U_0 = 0$, (i.e. the limit case $\sigma \to \infty$). From Eq. (4.2) we obtain

$$\bar{\psi}(\xi, y) = -\frac{y}{|y|}\frac{\exp\{-|y|(\xi^2 + k^2)^{\frac{1}{2}}\} - \exp\{-|y|(\xi^2 + \beta^2)^{\frac{1}{2}}\}}{(\beta^2 - k^2)}\frac{\bar{f}_{10}(\xi)}{2}, \tag{4.3}$$

and finally

$$\frac{\bar{f}_{10}(\xi)}{2} = \frac{N_+(i\alpha)N_-(\xi)}{i(\xi - i\alpha)}, \tag{4.4}$$

where $N(\xi) = (\xi^2 + \beta^2)^{\frac{1}{2}} + (\xi^2 + k^2)^{\frac{1}{2}}$ is analytic in the strip $-k < \mathrm{Im}(\xi) < k$.

Before we proceed we must determine $N_-(\xi)$. In the previous section we

* This result can be obtained from Eq. (21) of [1]. It also was given explicitly by Lagerstrom, Cole and Trilling [4].

could determine $K_-(\xi)$ by inspection; here the situation is more difficult. The details of the computation of $N_-(\xi)$ are presented in Appendix A. It is found that

$$N_-(\xi) = \exp\left\{-2\overline{\pi i}\int_0^{\xi} F(\xi)\,d\xi\right\}, \tag{4.5}$$

where in the limit as $k \to 0$

$$F(\xi) = \frac{1}{(\xi^2+\beta^2)^{\frac{1}{2}}} \ln \frac{\beta + (\xi^2+\beta^2)^{\frac{1}{2}}}{\beta - (\xi^2+\beta^2)^{\frac{1}{2}}}. \tag{4.6}$$

The behaviour of $F(\xi)$ in the complex plane is quite complicated. The points $\xi = i\beta$ and $\xi = 0$ are branch points of the function; however the behaviour of the logarithmic term in (4.6) must be such that $\xi = -i\beta$ is not a branch point. (Recall $N_-(\xi)$ must be analytic in the LHP.) Thus we choose the branch of the logarithm by defining the logarithmic term in (4.6) to be zero at $\xi = -i\beta$, in which case it becomes $-2\pi i$ at $\xi = +i\beta$. This can be seen by writing $F(\xi) = 2(\xi^2+\beta^2)^{\frac{1}{2}} \ln\{[\beta + (\xi^2+\beta^2)^{\frac{1}{2}}]/e^{\pi i/2}\xi\}$ where we have taken $(-1) = e^{\pi i}$. As our branch cut we shall take the entire positive imaginary axis. The determination of ξ and $(\xi^2+\beta^2)^{\frac{1}{2}}$ on the branch cuts is given in Appendix B where the results are used in evaluating $f_{10}(x)$. Finally from Eq. (4.5) it is clear that $N_-(0) = 1$, and since when $k \to 0$, $N(0) = \beta$ we see that $N_+(0) = \beta$.

We may now compute $f_{10}(x)$. Upon substituting in Eq. (4.4) and using the usual inversion formula we have, letting $k \to 0$, $\alpha \to 0$ that

$$\tfrac{1}{2} f_{10}(x) = \frac{\beta}{2\pi}\int_{-\infty}^{\infty} \frac{e^{i\xi x}}{i\xi} \exp\left\{-\frac{1}{2\pi i}\int_0^{\xi} F(\xi)\,d\xi\right\}, \tag{4.7}$$

where the path of integration must pass below the origin. This integral is evaluated in Appendix B. We obtain

$$\tfrac{1}{2} f_{10}(x) = \frac{\beta}{\pi}\left\{\pi + \int_0^{x'} e^{-y x'} H(y)\,dy + \int_1^{\infty} e^{-y x'} \frac{[y + (y^2+1)^{\frac{1}{2}}]^{\frac{1}{2}}}{1} P(y)\,dy\right\}. \tag{4.8}$$

Here $x' = \beta x = (1+i)(\omega/2\nu)^{\frac{1}{2}}x$,

$$H(y) = \exp\left\{-\frac{1}{\pi}\left[t \ln \cot\frac{t}{2} + \int_0^t \frac{\theta}{\sin\theta}\,d\theta\right]\right\}, \qquad t = \arcsin y,$$

$$P(y) = \exp\left\{-\frac{1}{\pi}\int_0^{m} \frac{\theta}{\sin\theta}\,d\theta\right\}, \qquad m = \arcsin(1/y),$$

and $0 < y < 1$ implies $0 < t < \pi/2$; also $1 < y < \infty$ implies $\pi/2 > m > 0$.

It is interesting to note, using Eq. (4.8) that $f_{10}(x)$ is of the form β times a function of βx. The expression for $f_{10}(x)$ given by Eq. (4.8) may be used for numerical computation when $(\omega/2\nu)^{\frac{1}{2}}x$ is neither large nor small. Of course the second integral must be converted into an integration over a finite range but this is not a difficult matter. Using Eq. (4.8), $f_{10}(x)/2$ has been computed for $hx = \frac{1}{2}$ and 1, where $h = (\omega/2\nu)^{\frac{1}{2}}$. The results of this computation are given at the end of this section. When $x = 0$ it is clear that the second integral in (4.8) does not converge; this of course is not surprising since we could anticipate

from the result of Sec. 3 that $f_{10}(x)$ should behave like $x^{-\frac{1}{2}}$ near the leading edge of the plate.

To complete our investigation of this case we shall obtain expressions for the behaviour of $f_{10}(x)$ for large and small values of hx. To do this we first note, since $f_{10}(x)$ is identically zero for $x < 0$, that the Fourier transform of $f_{10}(x)$ is essentially the usual Laplace transform but rotated through 90°. Hence we may make use of the Tauberian theorems relating the behaviour of the Laplace inverse at the origin to the behaviour of the function at infinity, and the behaviour of the Laplace inverse at infinity to the behaviour of the function at the origin. (See [5].) Thus we must study the behaviour of $N_-(\xi)$ for large ξ and small ξ.

The expression for $N_-(\xi)$ given by Eq. (4.5) can be written in the more convenient form

$$N_-(\xi) = \{s + (1 + s^2)^{\frac{1}{2}}\}^{\frac{1}{2}} e^{\pi i/4} \exp\left\{-\frac{i}{\pi}\int_0^\infty \ln(\coth\theta/2)\, d\theta\right\}, \tag{4.9}$$

where $s = \xi/\beta$ and $\theta = \sinh^{-1} s$. This result is obtained by rationalizing $\{\beta + (\xi^2 + \beta^2)^{\frac{1}{2}}\}/\{\beta - (\xi^2 + \beta^2)^{\frac{1}{2}}\}$ in (4.6), and then making the substitution $s = \sinh\theta$. It is not difficult to show from (4.9) that as $s \to 0$

$$N_-(\xi) \to 1, \tag{4.10a}$$

and that as $s \to \infty$

$$N_-(\xi) = \left\{\left(\frac{2\xi}{\beta}\right)^{\frac{1}{2}} - \frac{e^{\pi i/2}}{\pi}\left(\frac{2\beta}{\xi}\right)^{\frac{1}{2}}\right\} e^{\pi i/4} + O[(\beta/\xi)^{\frac{1}{2}}]. \tag{4.10b}$$

Using these results and Eq. (4.4) we obtain that as $x \to \infty$

$$\frac{1}{2} f_{10}(x) \to \beta = \left(\frac{\omega}{\nu}\right)^{\frac{1}{2}} e^{\pi i/4}, \tag{4.11a}$$

and as $x \to 0$

$$\frac{1}{2} f_{10}(x) = \left(\frac{2\beta}{\pi x}\right)^{\frac{1}{2}} \left\{1 + \frac{2\beta x}{\pi} + O[(\beta x)^2]\right\}. \tag{4.11b}$$

In particular for very small βx

$$\frac{1}{2} f_{10}(x) \to \left\{\left(\frac{\omega}{\nu}\right)^{\frac{1}{2}} \frac{2}{\pi x}\right\}^{\frac{1}{2}} e^{3\pi i/8}.$$

As could have been anticipated Eq. (4.11a) is the result for an infinite lamina undergoing simple harmonic motion. (See [6].) This result represents a wave of transversal vibration, i.e. a shear wave. Exactly how far down the plate it is necessary to go before (4.11a) is valid can be obtained by evaluating $f_{10}(x)$ from Eq. (4.8). The argument and a dimensionless modulus of $f_{10}(x)$ have been computed using Eqs. (4.8) and (4.11b). The results are tabulated in Table I, and are exhibited graphically in Figs. 3 and 4. It appears from Fig. 3 that Eq. (4.11a) is an excellent representation for $f_{10}(x)$ for $hx > 2.5$ approximately. Also from

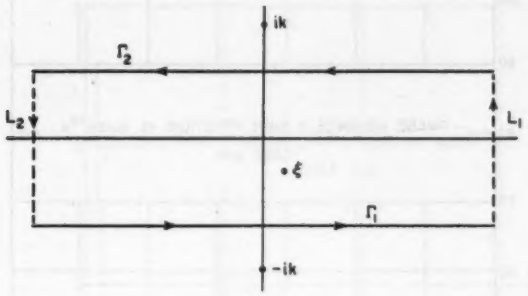


FIG. 1

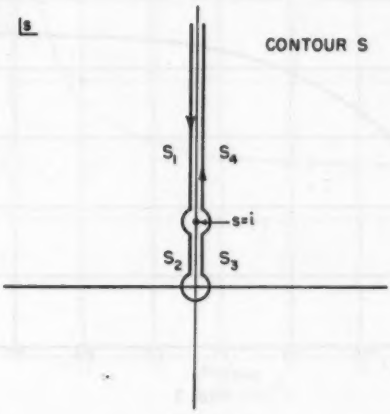


FIG. 2

TABLE I

$(\omega/2\nu)^{\frac{1}{2}}x$	Arg $[f_{10}(x)]$	$\lvert f_{10}(x)\rvert/2(\omega/2\nu)^{\frac{1}{2}}$	Evaluated by
$\to 0$	22.5°	$\to\infty$	
.1	25.93°	3.19	Asymptotic Formula Eq. (4.11b)
.25	29.97°	2.20	Asymptotic Formula Eq. (4.11b)
.50	36°	1.77	Asymptotic Formula Eq. (4.11b)
.50	34.65°	1.80	Numerical Computation Eq. (4.8)
1.0	39.15°	1.58	Numerical Computation Eq. (4.8)
$\to\infty$	45°	2[1]	

the tabulated results it appears that Eq. (4.11b) is valid for $hx \leqq \frac{1}{2}$ approximately.

To summarize our results we have, under the assumptions $(\omega\nu)^{\frac{1}{2}} \gg U_0$ and $(\omega\nu)^{\frac{1}{2}} \gg U_1$, that the skin friction τ is

$$\tau = \tfrac{1}{2}\mu U_0 f_0(x) + \mu U_1\{\tfrac{1}{2}f_{10}(x) + O(1/\sigma)\}e^{i\omega t} \tag{4.12}$$

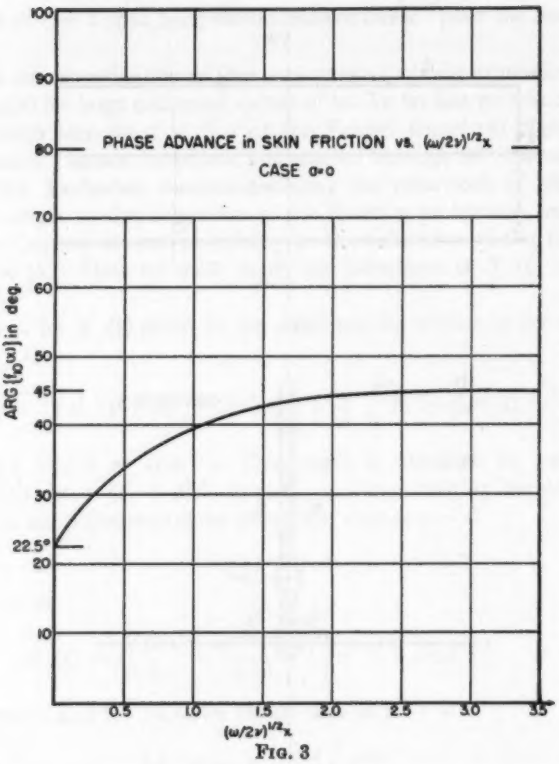


FIG. 3

where $f_0(x)/2$ is given by Eq. (3.11) and $f_{10}(x)/2$ is given by Eq (4.8). Near the leading edge of the plate, τ takes the form

$$\tau = \mu U_0 \left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} + \mu U_1 \left\{\left(\frac{\omega}{\nu}\right)^{\frac{1}{2}} \left(\frac{2}{\pi x}\right)\right\}^{\frac{1}{2}} e^{i(\omega t+\pi/8)}. \tag{4.13}$$

Far down the plate, i.e. $hx > 2.5$ approximately,

$$\tau = \mu U_0 \left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} + \mu U_1 \left(\frac{\omega}{\nu}\right)^{\frac{1}{2}} e^{i(\omega t+\pi/4)}. \tag{4.14}$$

Thus the flow consists of the sum of a mean flow plus an oscillatory motion which is independent of the mean motion. The phase advance in the skin friction increases monotonically from $\pi/8$ at the leading edge to $\pi/4$ in a distance of $hx = 2.5$ approximately.

5. Case $U_0 + U_1 e^{i\omega t}$. Now we are concerned with solving Eq. (2.6) without any restriction on the parameters involved. In theory we could, of course, proceed exactly as we have in the past two sections. However because of the nonhomogeneous term as well as the more complicated differential operator which is

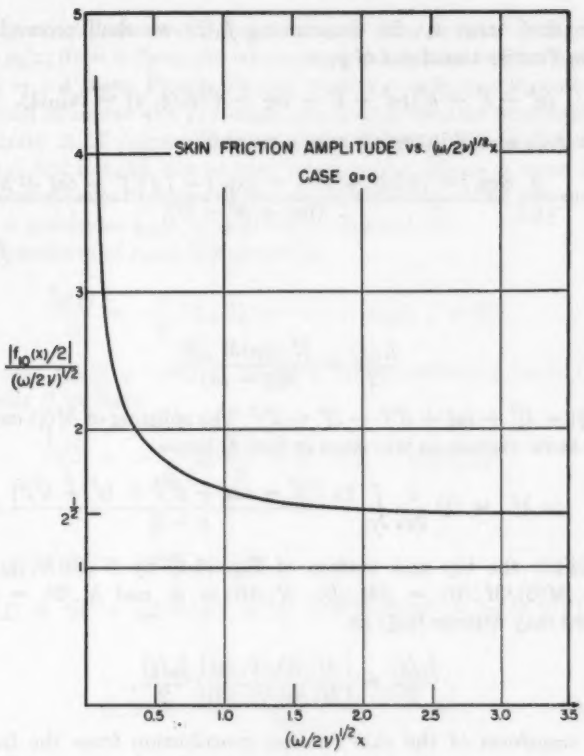


FIG. 4

applied to ψ we should expect the details of the computation to be more difficult. In actual fact a direct application of the technique used in the previous sections does present some rather formidable problems; hence it appears desirable to attack this problem in a slightly different manner.

First we note that ψ can be expressed as the sum of an appropriate solution of the homogeneous equation plus a solution of the non-homogeneous equation. Hence we can write

$$\psi(x, y) = \psi_1(x, y) + \psi_2(x, y), \tag{5.1}$$

where ψ_1 is a solution of the homogeneous equation which also satisfies the boundary condition $\partial\psi_1/\partial y = -1$ on the plate. Then the solution, ψ_2, of the non-homogeneous equation (2.6), must satisfy the condition $\partial\psi_2/\partial y = 0$ on the plate. The discontinuity condition in ψ_{yy} can be expressed as $f_1(x) = f_{11}(x) + f_{12}(x)$ where

$$\psi_{1yy}(x, 0+) = -\psi_{1yy}(x, 0-) = \tfrac{1}{2}f_{11}(x),$$
$$\psi_{2yy}(x, 0+) = -\psi_{2yy}(x, 0-) = \tfrac{1}{2}f_{12}(x).$$

First we shall treat ψ_1. In determining $f_{11}(x)$ we shall proceed exactly as before. The Fourier transform of ψ_1 is

$$(\eta^2 + \xi^2 + k^2)\{\eta^2 + \xi^2 + ia\xi + \beta^2\}\psi_1(\xi, \eta) = i\eta \bar{f}_{11}(\xi). \tag{5.2}$$

Solving for $\psi_1(\xi, \eta)$ and inverting over η we obtain

$$\psi_1(\xi, y) = -\frac{y}{|y|} \frac{\exp\{-|y|(\xi^2 + k^2)\} - \exp\{-|y|(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}}{(ia\xi + \beta^2 - k^2)} \frac{\bar{f}_{11}(\xi)}{2}. \tag{5.3}$$

Then

$$\frac{\bar{f}_{11}(\xi)}{2} = \frac{M_+(ia)M_-(\xi)}{i(\xi - ia)}, \tag{5.4}$$

where $M(\xi) = (\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}} + (\xi^2 + k^2)^{\frac{1}{2}}$. The splitting of $M(\xi)$ can be carried out in the same manner as was done in Sec. 4, hence

$$\ln M_- = (\xi) \frac{1}{2\pi i} \int_{\Gamma_2} \frac{\ln\{(z^2 + iaz + \beta^2)^{\frac{1}{2}} + (z^2 + k^2)^{\frac{1}{2}}\}}{z - \xi} dz. \tag{5.5}$$

If we multiply the top and bottom of Eq. (5.4) by $N_+(0)N_-(\xi)$, note that $M_+(0) = M(0)/M_-(0) = \beta M_-(0)$, $N_+(0) = \beta$, and $N_-(0) = 1$, and use Eq. (4.4) we may express $\bar{f}_{11}(\xi)$ as

$$\frac{\bar{f}_{11}(\xi)}{2} = \left\{\frac{M_-(\xi)/N_-(\xi)}{M_-(0)/N_-(0)}\right\} \frac{\bar{f}_{10}(\xi)}{2}. \tag{5.6}$$

Thus the transform of the skin friction contribution from the homogeneous solution can be expressed as a function of ξ times the transform of the skin friction when $U_0 = 0$. Equation (5.6) will be examined in more detail shortly, but now let us turn our attention to determining the non-homogeneous solution of Eq. (2.6).

Using Eq. (2.5) it is clear that $-(a/\beta^2)\chi_x$ is a solution of Eq. (2.6). However an inspection of the properties of χ as given in Sec. 3 shows that for $y = 0$, $\psi_2 = -(a/\beta^2)\chi_x$ will lead to a velocity which becomes infinite at the leading edge of the plate, i.e. $\chi_{xy}(x, 0)$ behaves like $x^{-\frac{1}{2}}$ for small $x > 0$. This, of course, we can not accept. But we can add solutions of the homogeneous equation to this solution, and the result will still be a solution of the non-homogeneous equation. Hence we might be able to construct an appropriate non-homogeneous solution by finding a homogeneous solution which will destroy the undesirable singularity in χ_x. Thus we take ψ_2 to be of the form

$$\psi_2(x, y) = -\frac{a}{\beta^2}\{\chi_x(x, y) + \phi(x, y)\}, \tag{5.7}$$

where $\phi(x, y)$ must satisfy the homogeneous equation $(\Delta - a\partial/\partial x - \beta^2)\Delta\phi = 0$ and also remove the singularity in χ_{xy}. To describe our problem completely

let us specify exactly the conditions which we must require of ψ_2. The first condition is $\psi_2(x, 0) = 0$. Secondly we require that $\partial\psi_2/\partial y = 0$ on $y = 0, x > 0$, since $\partial\psi_1/\partial y = -1$ there. Finally $\partial^2\psi_2/\partial y^2 = 0$ on $y = 0, x < 0$, since $f_1(x) \equiv 0$ for $x < 0$. And of course also ψ_2's behaviour is such that its contribution to the velocity is finite at all points. The first three of these requirements are clearly satisfied by χ_x but not the last as mentioned earlier. Hence ϕ must satisfy the first three requirements and also fix the singularity in χ_x. The determination of $\phi(x, y)$ and in particular $\bar{\phi}_{xy}(\xi, 0)$ is given in Appendix C.

Now the transform of $f_{12}(x)/2$ is given by

$$\frac{\bar{f}_{12}(\xi)}{2} = -\frac{a}{\beta^2}\{\bar{\chi}_{xyy}(\xi, y = 0) + \phi_{yy}(\xi, y = 0)\}, \tag{5.8}$$

and noting that $\bar{\chi}_{xyy}(\xi, y = 0) = i\xi\bar{f}_0(\xi)/2$, and substituting for $\bar{\phi}_{yy}(\xi, y = 0)$ from Appendix C we have

$$\frac{\bar{f}_{12}(\xi)}{2} = -\frac{a}{\beta^2}\left\{i\xi\frac{\bar{f}_0(\xi)}{2} - e^{\pi i/4}a^{\frac{1}{2}}(\xi - ia_1)^{\frac{1}{2}}Q_-(\xi)\right\}, \tag{5.9}$$

where

$$a_1 = \{(a^2 + 4\beta^2)^{\frac{1}{2}} - a\}/2,$$

$$Q(\xi) = \{(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}} + (\xi^2 + k^2)^{\frac{1}{2}}\}/\{2(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}\}$$

and

$$Q_-(\xi) = \exp\left\{-\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{\ln Q(x)}{x - \xi}\,dx\right\}, \qquad \text{Im}(\xi) < 0.$$

Combining Eqs. (5.9) and (5.7) we see that

$$\frac{\bar{f}_1(\xi)}{2} = \left\{\frac{M_-(\xi)/N_-(\xi)}{M_-(0)/N_-(0)}\right\}\frac{\bar{f}_{10}(\xi)}{2} - \frac{a^2}{\beta}\left\{i\xi\frac{\bar{f}_0(\xi)}{2} - e^{\pi i/4}a^{\frac{1}{2}}(\xi - ia_1)^{\frac{1}{2}}Q_-(\xi)\right\}. \tag{5.10}$$

We might recall in considering this equation that $\bar{f}_0(\xi)/2$ is the skin friction when $U_1 = 0$, and $\bar{f}_{10}(\xi)/2$ is the skin friction when $U_0 = 0$, $(\omega\nu)^{\frac{1}{2}} > U_1$.

To invert Eq. (5.10) in general is a rather difficult task; however it is possible to obtain expressions for $f_1(x)$ for small x and for large x. First let us consider $\bar{f}_{11}(\xi)/2$. It is convenient to use the expression for $\bar{f}_{11}(\xi)/2$ given in Eq. (5.6). As $\xi \to 0$, (i.e. $x \to \infty$) the bracketed quantity in Eq. (5.6) is invariant with changes in a, and as a matter of fact approaches the value one. Here the first approximation to $f_{11}(x)/2$ for large x is simply

$$\tfrac{1}{2}f_{11}(x) \sim \tfrac{1}{2}f_{10}(x) = (\omega/\nu)^{\frac{1}{2}}e^{\pi i/4} \qquad \text{as } x \to \infty. \tag{5.11}$$

On the other hand $M_-(\xi)/N_-(\xi) \to 1$ as $\xi \to \infty$ and hence we have

$$\frac{f_{11}(x)}{2} \sim \frac{N_-(0)}{M_-(0)}\frac{f_{10}(x)}{2}, \qquad \sim \frac{N_-(0)}{M_-(0)}\left\{\frac{2}{\pi x}\left(\frac{\omega}{\nu}\right)^{\frac{1}{2}}\right\}^{\frac{1}{2}}e^{\pi i/8} \qquad \text{as } x \to 0. \tag{5.12}$$

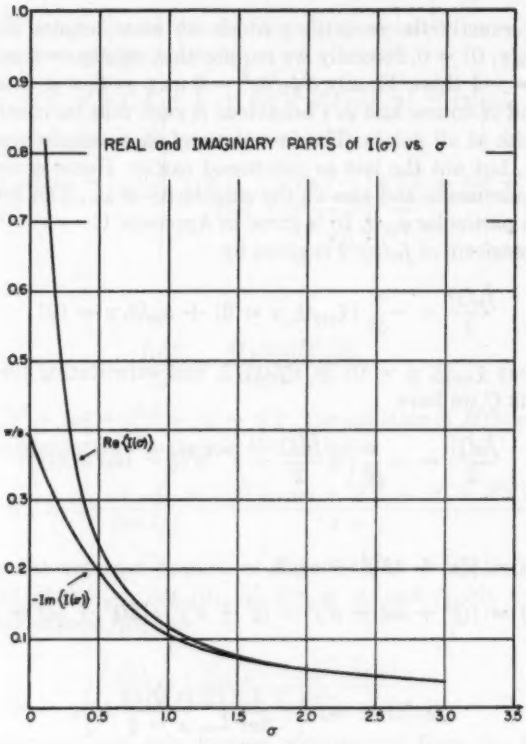


FIG. 5

The ratio $N_-(0)/M_-(0)$ is a function of the parameter $\sigma^2 = \omega\nu/(cU_0)^2$. It is computed in Appendix D where we obtain upon letting $N_-(0)/M_-(0) = \exp\{I(\sigma)\}$ that for $\sigma \geqq \frac{1}{2}$

$$I(\sigma) = -\frac{i}{\pi}\sum_{n=0}^{\infty}\frac{(\gamma/\sigma)^{2n+1}}{(2n+1)^2},$$

and for $\sigma \leqq \frac{1}{2}$

$$I(\sigma) = -\frac{\pi i}{8} - \frac{1}{2}\ln(2\sigma) + \frac{i}{\pi}\sum_{n=0}^{\infty}\frac{(\sigma/\gamma)^{2n+1}}{(2n+1)^2},$$

where $\gamma = 2^{-1}e^{\pi i/4}$. From these expressions for $I(\sigma)$ we can already see the effect of the mean flow on the oscillatory motion. For instance as $U_0 \to \infty$, (i.e. $\sigma \to 0$), $I(\sigma) \to \{-\pi i/8 - 2^{-1}\ln(2\sigma)\}$ and upon substituting in Eq. (5.12) we see that the phase advanced in $f_{11}(x)/2$ is cancelled. We shall discuss this in more detail after we have obtained appropriate representations for $f_{12}(x)$. The real and imaginary parts of $I(\sigma)$ are given graphically in Fig. 5.

Now $\hat{f}_{12}(\xi)$ is given by Eq. (5.9). To determine $f_{12}(x)$ for small x we must know

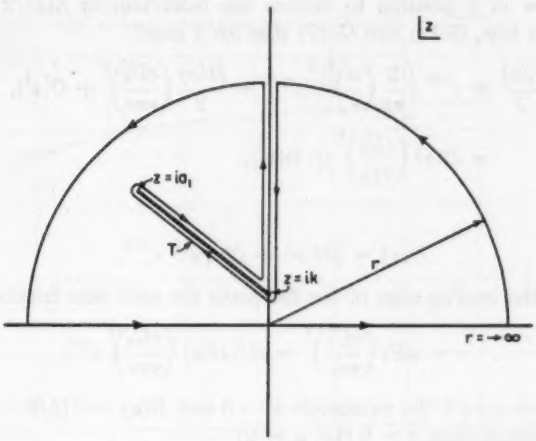


FIG. 6

the behaviour of $\hat{f}_{12}(\xi)$ as $\xi \to \infty$. In particular we must study $Q_-(\xi)$ for ξ large. This is done in Appendix E where we obtain the result that

$$Q_-(\xi) = 1 - \frac{G_1}{2\pi i}\frac{1}{\xi} + O\left(\frac{1}{\xi^2}\right) \tag{5.13}$$

where

$$G_1 = -ia_1\left\{-\frac{\delta}{1-\delta}\pi i + 1 + \left(\frac{2}{1-\delta} - \frac{1-\delta}{2i\delta^{\frac{1}{2}}}\right)\log\frac{1+i\delta^{\frac{1}{2}}}{1-i\delta^{\frac{1}{2}}}\right\}, \tag{5.14}$$

and $\delta = a_1/a_2$, $a_2 = \{(a^2 + 4\beta^2)^{\frac{1}{2}} + a\}/2$. Upon substituting for $Q_-(\xi)$ and $f_0(\xi)$ in Eq. (5.9) we obtain that for large ξ

$$\frac{\hat{f}_{12}(\xi)}{2} = \frac{\dot{a}^{\frac{1}{2}}a_1}{2\beta^2}\left\{1 + \frac{e^{\pi i/2}G'}{\pi}\right\}\frac{1}{(i\xi)^{\frac{1}{2}}} + O(\xi^{-\frac{1}{2}}), \tag{5.15}$$

where G' is the bracketed quantity in Eq. (5.14). Finally using the definitions of a_1, a_2, δ, β^2 and σ Eq. (5.15) can be written as

$$\frac{\hat{f}_{12}(\xi)}{2} = \frac{H(\sigma)}{2}\left(\frac{a}{i\xi}\right)^{\frac{1}{2}} + O(\xi^{-\frac{1}{2}}), \tag{5.16}$$

where

$$H(\sigma) = 1 + \frac{1}{\pi}\left\{\left[2i - \frac{(1-\delta)^2}{2\delta^{\frac{1}{2}}}\right]\log\frac{1+i\delta^{\frac{1}{2}}}{1-i\delta^{\frac{1}{2}}} + i(1-\delta)\right\}$$

and $\delta = 1 + i\{(1 + 4i\sigma^2)^{\frac{1}{2}} - 1\}/2\sigma^2$. Consequently

$$\frac{f_{12}(x)}{2} = \frac{H(\sigma)}{2}\left(\frac{cU_0}{\nu\pi x}\right)^{\frac{1}{2}} + O(x^{\frac{1}{2}}) \qquad \text{as } x \to 0. \tag{5.17}$$

We are now in a position to discuss the behaviour of $f_1(x)/2$ for small x. We have from Eqs. (5.12) and (5.17) that for x small

$$\begin{aligned}\frac{f_1(x)}{2} &= e^{I(\sigma)}\left\{\frac{2}{\pi x}\left(\frac{\omega}{\nu}\right)^{\frac{1}{2}}\right\}^{\frac{1}{2}} e^{\pi i/8} + \frac{H(\sigma)}{2}\left(\frac{cU_0}{\nu\pi x}\right)^{\frac{1}{2}} + O(x^{\frac{1}{2}}),\\ &= D(\sigma)\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} + O(x^{\frac{1}{2}}),\end{aligned} \tag{5.18}$$

where

$$D(\sigma) = \tfrac{1}{2}H(\sigma) + (2\sigma)^{\frac{1}{2}}e^{I(\sigma)}e^{\pi i/8}.$$

Thus near the leading edge of the flat plate the total skin friction is

$$\tau = \mu U_0\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} + \mu U_1 D(\sigma)\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} e^{it}. \tag{5.19}$$

As $U_0 \to \infty$, or $\omega \to 0$ the parameter $\sigma \to 0$ and $D(\sigma) \to \{(3/2 - 7\sigma^2/12\pi) + i4\sigma^2/\pi^2\}$ and hence when $\sigma = 0$ (i.e. $\omega = 0$)

$$\tau = \mu U_0\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}}\left(1 + \frac{3}{2}\frac{U_1}{U_0}\right). \tag{5.20}$$

This result is in agreement with Eq. (3.13). As a matter of fact it will be shown, at the end of this section, using Eq. (5.10) that the above result is valid for all x when $\sigma = 0$.

On the other hand when $\sigma \to \infty$, i.e. $\omega \to \infty$ we have

$$D(\sigma) \to (2\sigma)^{\frac{1}{2}}\{1 + (1 - i)/(2\pi\sigma)\}e^{\pi i/8}$$

and hence

$$\tau = \mu U_0\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}} + \mu U_1\left\{\frac{2}{\pi x}\left(\frac{\omega}{\nu}\right)\right\}^{\frac{1}{2}} e^{i(\omega t + \pi/8)}. \tag{5.21}$$

This is in agreement with Eq. (4.12) as could have been anticipated. As σ increases from 0 to ∞, arg $\{D(\sigma)\}$ and hence the phase advance in the fluctuating skin at the leading edge increases monotonically from 0 to $\pi/8$. This result is presented graphically in Fig. 7. Also the modulus of $D(\sigma)$ is given in Fig. 8. It can be seen from Fig. 8 that the modulus decreases for a short range of σ as σ increases. This decrease is extremely slight, of the order of two or three in one hundred and fifty. However the asymptotic formula for $D(\sigma)$ for small σ bears out the numerical calculations. There is apparently no obvious physical reason why the amplitude of the skin friction should decrease slightly as σ increases from 0 to roughly $\frac{1}{2}$.

It remains now to consider the large x behaviour of $f_{12}(x)$. (We have already shown that $f_{11}(x)/2 \to (\omega/\nu)e^{\pi i/4}$ as $x \to \infty$, see Eq. (5.11), So we must now consider the small ξ behaviour of $\tilde{f}_{12}(\xi)$. This is a fairly laborious task and is perhaps not worth a detailed analysis. It can be seen from the definition of $Q_-(\xi)$ that $Q_-(0)$ is finite and hence $(\xi - ia_1)^{\frac{1}{2}}Q_-(\xi)$ behaves like $\xi^{\frac{1}{2}}$ for small ξ at worst. Also $i\xi\tilde{f}_0(\xi)/2$ behaves like $(i\xi)^{\frac{1}{2}}$ so $\tilde{f}_{12}(\xi)/2$, (see Eq. 5.9), will consist

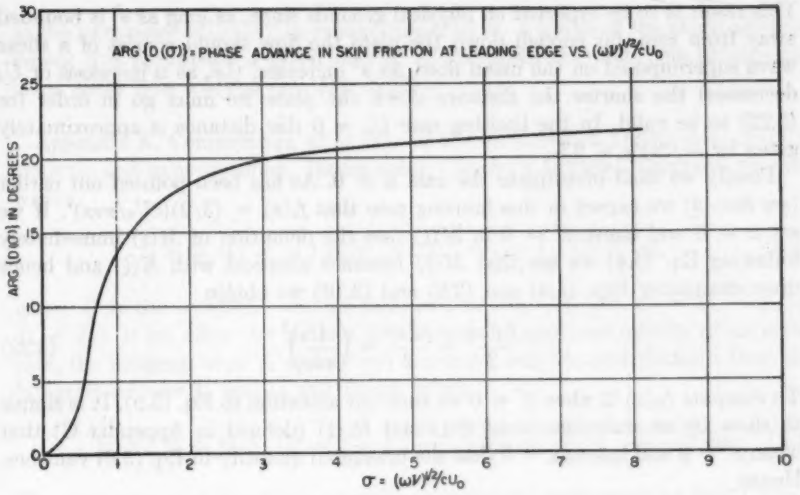


FIG. 7

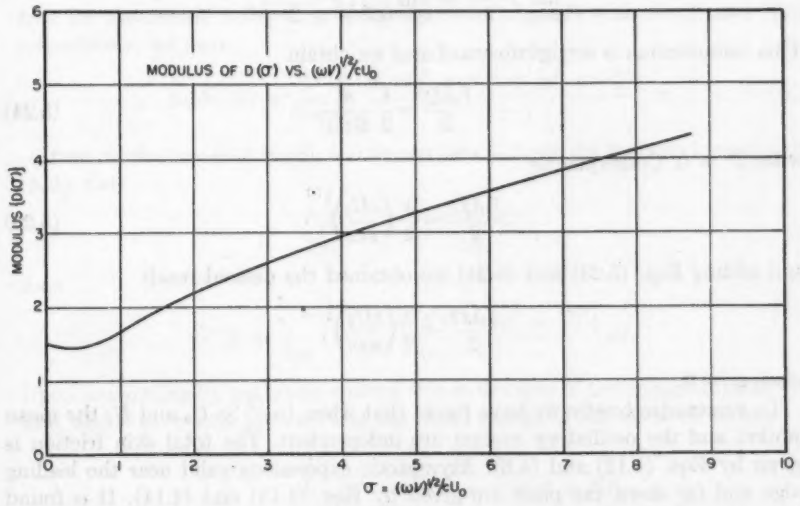


FIG. 8

of terms of the type $(i\xi)^{\frac{1}{2}}$, $(i\xi)^{\frac{1}{4}}$, and so on, at worst. Hence $f_{11}(x)/2$ will be made up of terms like $x^{-\frac{1}{2}}$, $x^{-\frac{1}{4}}$, for x large. Consequently as long as $a_1 \neq 0$, (i.e. $\omega \neq 0$) these terms should die out more quickly than those contributed by $f_{11}(x)$; hence

$$f_1(x) \to \left(\frac{\omega}{\nu}\right) e^{\pi i/4}, \qquad \text{as } x \to \infty \quad (5.22)$$

This result is to be expected on physical grounds since, as long as σ^2 is bounded away from zero, far enough down the plate the flow should consist of a shear wave superimposed on the mean flow. As σ^2 increases, (i.e. as ω increases or U_0 decreases) the shorter the distance down the plate we must go in order for (5.22) to be valid. In the limiting case $U_0 = 0$ this distance is approximately given by $(\omega/2\nu)^{\frac{1}{2}}x = 2.5$.

Finally we shall investigate the case $\omega = 0$. As has been pointed out earlier (see Sec. 3) we expect in this limiting case that $f_1(x) = (3/2)(cU_0/\pi\nu x)^{\frac{1}{2}}$. If we set $\omega = 0$ and hence $\beta^2 = 0$ in $M(\xi)$ (see the definition of $M(\xi)$ immediately following Eq. (5.4) we see that $M(\xi)$ becomes identical with $K(\xi)$ and hence upon examining Eqs. (5.4) and (3.9) and (3.10) we obtain

$$\frac{f_{11}(x)}{2} = \frac{f_0(x)}{2} = \left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}}. \tag{5.23}$$

To compute $f_{12}(x)/2$ when $\beta^2 = 0$ we turn our attention to Eq. (5.9). It is simple to show by an examination of $Q(\xi)$ and $R_-(\xi)$ (defined in Appendix C) that when $\beta^2 = 0$ and hence $a_1 = 0$ that the bracketed quantity in Eq. (5.9) vanishes. Hence

$$\lim_{\beta^2 \to 0} \frac{\bar{f}_{12}(\xi)}{2} = \lim_{\beta^2 \to 0} \frac{d}{d\beta^2}\left\{\beta^2 \frac{\bar{f}_{12}(\xi)}{2}\right\}.$$

This computation is straightforward and we obtain

$$\frac{\bar{f}_{12}(\xi)}{2} = \frac{1}{2}\frac{a^{\frac{1}{2}}}{2(i\xi)^{\frac{1}{2}}} \tag{5.24}$$

when $\beta^2 = 0$. Consequently

$$\frac{f_{12}(x)}{2} = \frac{1}{2}\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}}, \tag{5.25}$$

and adding Eqs. (5.23) and (5.24) we obtained the desired result

$$\frac{f_1(x)}{2} = \frac{3}{2}\left(\frac{cU_0}{\pi\nu x}\right)^{\frac{1}{2}},$$

when $\omega = 0$.

To summarize briefly we have found that when $(\omega\nu)^{\frac{1}{2}} \gg U_0$ and U_1 the mean motion and the oscillatory motion are independent. The total skin friction is given by Eqs. (4.12) and (4.8). Asymptotic expressions valid near the leading edge and far down the plate are given by Eqs. (4.13) and (4.14). It is found that the phase advance in the time dependent skin friction increases from $\pi/8$ at the leading edge to $\pi/4$ in a distance of approximately $(\omega/2\nu)^{\frac{1}{2}}x = 2.5$. Finally when the interaction of the mean flow and the time dependent flow can not be neglected a result valid near the leading edge and one valid far downstream are obtained for the oscillating skin friction. These results are given in Eqs. (5.19), (5.20), (5.21) and (5.22). It is found that the phase advance in the time dependent skin friction at the leading edge increases from 0 to $\pi/8$

as the parameter $cU_0/(\omega\nu)^2$ increases from 0 to ∞. A graph of the phase advance versus this parameter is given in Fig. 7. Far down the plate one obtains the usual shear flow solution (see Eq. (5.22)).

Appendix A. Computation of $N_-(\xi)$. In the strip $-k < \mathrm{Im}(\xi) < k$, $N(\xi)$ is analytic and non-vanishing. Hence using Cauchy's Integral Theorem we may write

$$\ln N(\xi) = \ln \{N_+(\xi) \cdot N_-(\xi)\} = \frac{1}{2\pi i} \int_C \frac{\ln N(\xi)}{z - \xi}\, dz, \tag{A.1}$$

where C is the complete contour indicated in Fig. (1), (i.e. $C = \Gamma_1 + L_1 + \Gamma_2 + L_2$). If we allow the ends of this contour to approach infinity at an equal rate, the integrals over Γ_1 and Γ_2 will converge, and the contributions from the integrations over L_1 and L_2 will vanish giving

$$\ln N_+(\xi) + \ln N_-(\xi) = \frac{1}{2\pi i} \int_{\Gamma_1} \frac{\ln N(z)\, dz}{z - \xi} + \frac{1}{2\pi i} \int_{\Gamma_2} \frac{\ln N(\xi)}{z - \xi}\, dz. \tag{A.2}$$

The first integral in the above equation is uniformly convergent for $\mathrm{Im}(\xi) > -k$ and hence represents an analytic function of ξ there. A similar statement is true for the second integral. Associating these integrals with $N_+(\xi)$ and $N_-(\xi)$ respectively, we have

$$\ln N_-(\xi) = \frac{1}{2\pi i} \int_{\Gamma_2} \frac{\ln \{(z^2 + \beta^2)^{\frac{1}{2}} + (z^2 + k^2)^{\frac{1}{2}}\}}{z - \xi}\, dz. \tag{A.3}$$

If we confine ξ so that $\mathrm{Im}(\xi) < 0$ we can take Γ_2 to be the real axis in evaluating (A.3). Let

$$\frac{F(\xi)}{2\pi i} = -\frac{d\{\ln N_-(\xi)\}}{d\xi},$$

then

$$F(\xi) = \int_{-\infty}^{\infty} \frac{\ln \{(x^2 + \beta^2)^{\frac{1}{2}} + (x^2 + k^2)^{\frac{1}{2}}\}}{(x - \xi)^2}\, dx.$$

Upon integrating by parts and writing this as the sum of two integrals from $-\infty$ to 0, and 0 to ∞ and then making the usual change of variables we obtain

$$F(\xi) = \int_0^{\infty} \frac{2x^2\, dx}{(x^2 - \xi^2)(x^2 + k^2)^{\frac{1}{2}}(x^2 + \beta^2)^{\frac{1}{2}}} \tag{A.4}$$

In the limit as $k \to 0$ the value of this integral is

$$F(\xi) = \frac{1}{(\xi^2 + \beta^2)^{\frac{1}{2}}} \ln \frac{\beta + (\xi^2 + \beta^2)^{\frac{1}{2}}}{\beta - (\xi^2 + \beta^2)^{\frac{1}{2}}}. \tag{A.5}$$

Now we can specify $N_-(\xi)$ precisely by

$$N_-(\xi) = \exp\left\{-\frac{1}{2\pi i} \int_0^{\xi} F(\xi)\, d\xi\right\}. \tag{A.6}$$

The lower limit in the above integral could be chosen differently but this merely multiplies $N_-(\xi)$ by a constant; the reciprocal of this constant would then multiply $N_+(\xi)$.

Appendix B. Evaluation of $f_{10}(x)/2$. Substituting in Eq. (4.4) for $N_-(\xi)$, and using the usual inversion formula we have letting $k \to 0$, $\alpha \to 0$ that

$$\frac{1}{2}f_{10}(x) = \frac{\beta}{2\pi}\int_{-\infty}^{\infty}\frac{e^{i\xi x}}{i\xi}\exp\left\{-\frac{1}{2\pi i}\int_0^{\xi}F(\xi)\,d\xi\right\}d\xi, \tag{B.1}$$

where the path of integration must pass below the origin. Making the change of variables $s = \xi/\beta, x' = \beta x$ and noting that for $x > 0$ we can deform our path of integration to an integration over S (see Fig. 2) leads to

$$\frac{1}{2}f_{10}(x) = \frac{\beta}{2\pi}\int_S\frac{e^{isx'}}{is}G(s)\,ds, \tag{B.2}$$

where

$$\begin{aligned} G(s) &= \exp\left\{-\frac{1}{2\pi i}\int_0^s\frac{1}{(1+s^2)^{\frac{1}{2}}}\ln\frac{1+(1+s^2)^{\frac{1}{2}}}{1-(1+s^2)^{\frac{1}{2}}}\,ds\right\}, \\ &= \exp\left\{-\frac{1}{\pi i}\int_0^s\frac{1}{(1+s^2)^{\frac{1}{2}}}\ln\frac{1+(1+s^2)^{\frac{1}{2}}}{e^{i\pi/2}s}\,ds\right\}. \end{aligned} \tag{B.3}$$

As indicated in Fig. 2, the contour S is broken up into the four parts S_1, S_2, S_3, S_4, plus the enclosures about the branch points at $s = i$ and the origin. It is not difficult to show, that as the radii of these enclosures are allowed to approach zero, that there are no contributions due to the integrations around the branch point $s = i$, and that the integration around the origin contributes 2π. On the S_1 we have the following:

$S_1 : s = e^{-3\pi i/2}y$; for $y > 1$, $(1-y^2)^{\frac{1}{2}} = e^{-\pi i/2}(y^2-1)^{\frac{1}{2}}$

$$G(s) = -i\{y+(y^2-1)^{\frac{1}{2}}\}^{\frac{1}{2}}\exp\left\{+\frac{1}{2\pi i}\int_y^{\infty}\frac{1}{(y^2-1)^{\frac{1}{2}}}\ln\frac{(y^2-1)^{\frac{1}{2}}-i}{(y^2-1)^{\frac{1}{2}}+i}\,dy\right\},$$

$S_2 : s = e^{-3\pi i/2}y$, $y < 1$

$$G(s) = \exp\left\{-\frac{1}{\pi}\int_0^y\frac{\ln[1+(1-y^2)^{\frac{1}{2}}]-\ln y}{(1-y^2)^{\frac{1}{2}}}\,dy - i\arcsin y\right\},$$

$S_3 : s = e^{\pi i/2}y$, $y < 1$

$$G(s) = \exp\left\{-\frac{1}{\pi}\int_0^y\frac{\ln[1+(1-y^2)^{\frac{1}{2}}]-\ln y}{(1-y^2)^{\frac{1}{2}}}\,dy + i\arcsin y\right\},$$

$S_4 : s = e^{\pi i/2}y$, for $y > 1$, $(1-y^2)^{\frac{1}{2}} = e^{\pi i/2}(y^2-1)^{\frac{1}{2}}$

$$G(s) = i\{y+(y^2-1)^{\frac{1}{2}}\}^{\frac{1}{2}}\exp\left\{\frac{1}{2\pi i}\int_y^{\infty}\frac{1}{(y^2-1)^{\frac{1}{2}}}\ln\frac{(y^2-1)^{\frac{1}{2}}-i}{(y^2-1)^{\frac{1}{2}}+i}\,dy\right\}.$$

In determining these expressions for $G(s)$ it was found convenient on S_1 and S_4 to write the integral 0 to y as the sum of the integrals 0 to 1 and 1 to ∞ minus the integral y to ∞. Combining all of these results we obtain finally

$$\frac{1}{2}f_{10}(x) = \frac{\beta}{\pi}\left\{\pi + \int_0^{'} e^{-\nu s'}H(y)\,dy + \int_1^{\infty} e^{-\nu s'}\,\frac{[y + (y^2-1)^{\frac{1}{2}}]^{\frac{1}{2}}}{y}\,P(y)\,dy\right\}. \quad \text{(B.4)}$$

Here

$$H(y) = \exp\left\{-\frac{1}{\pi}\int_0^{y}\frac{\ln[1+(1-u^2)] - \ln u}{(1-u^2)^{\frac{1}{2}}}\,du\right\}, \qquad y < 1$$

$$= \exp\left\{-\frac{1}{\pi}\left[t\ln\cot\frac{t}{2} + \int_0^{t}\frac{\theta}{\sin\theta}\,d\theta\right]\right\}, \qquad t = \arcsin y,$$

and

$$P(y) = \exp\left\{\frac{1}{2\pi i}\int_y^{\infty}\frac{1}{(y^2-1)^{\frac{1}{2}}}\ln\frac{(y^2-1)^{\frac{1}{2}} - i}{(y^2-1)^{\frac{1}{2}} + i}\,dy\right\}, \qquad y > 1$$

$$= \exp\left\{-\frac{1}{\pi}\int_0^{m}\frac{\theta}{\sin\theta}\,d\theta\right\}, \qquad m = \arcsin\frac{1}{y}.$$

When y ranges from 0 to 1, t goes from 0 to $\pi/2$; similarly $1 < y < \infty$ implies $\pi/2 > m > 0$.

Appendix C. Computation of $\bar{\phi}(\xi, y)$. The x Fourier transform of a function $\phi(x, y)$ which satisfies the differential equation $(\Delta - a\partial/\partial x - \beta^2)\Delta\phi = 0$ and the condition $\phi(x, 0) = 0$ is

$$\bar{\phi}(\xi, y) = \frac{y}{|y|}\,\frac{\exp\{-|y|(\xi^2+k^2)^{\frac{1}{2}}\} - \exp\{-|y|(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}\}}{A(\xi)}, \quad \text{(C.1)}$$

where $A(\xi)$ is an arbitrary function of ξ. The condition that $\partial\psi_2/\partial y = 0$ on $y = 0$, $x > 0$ requires $\bar{\phi}_y(\xi, y = 0)$ be analytic in the UHP, and the condition $\partial^2\psi_2/\partial y^2 = 0$ on $y = 0$, $x < 0$ requires $\bar{\phi}_{yy}(\xi, y = 0)$ be analytic in the LHP. These conditions may be written as

$$\bar{\phi}_y(\xi, 0) = \frac{-(\xi^2+k^2)^{\frac{1}{2}} + (\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}}{A(\xi)} = C_+(\xi), \quad \text{(C.2)}$$

$$\bar{\phi}_{yy}(\xi, 0) = \frac{(\xi^2+k^2) - (\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}}{A(\xi)} = D_-(\xi), \quad \text{(C.3)}$$

where $C_+(\xi)$ and $D_-(\xi)$ are functions analytic in the UHP and LHP respectively. The final condition on ϕ may be expressed as $\chi_{2y}(x, 0) + \phi_y(x, 0)$ be finite at the origin. Since the behaviour of the transform at infinity determines the behaviour of the function at the origin (see Sec. 4 and [5]), and since differentiation with respect to x implies multiplication of the transform by $i\xi$ we may express this condition as

$$\bar{\psi}_{2y}(x, 0) = -\frac{a}{\beta^2}\{i\xi\bar{\chi}_y(x, 0) + \bar{\phi}_y(x, 0)\} \sim \xi^{-(1+\epsilon)},$$

as $\xi \to \infty$, where $\epsilon > 0$. Using Eqs. (3.5), (3.10), and (C.2) this becomes

$$\left\{\frac{e^{-3\pi i/4}a^{\frac{1}{2}}}{2\xi^{\frac{1}{2}}} + C_+(\xi)\right\} \sim \xi^{-(1+\epsilon)}, \tag{C.4}$$

as $\xi \to \infty$. Hence $C_+(\xi) \sim \xi^{-\frac{1}{2}} + O[\xi^{-(1+\epsilon)}]$ as $\xi \to \infty$.

We can determine the functions $C_+(\xi)$ and $D_-(\xi)$ by an application of the Wiener-Hopf technique. First dividing (C.3) by (C.2) gives

$$D_-(\xi)/C_+(\xi) = -\{(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}} + (\xi^2 + k^2)^{\frac{1}{2}}\} = -M(\xi), \tag{C.5}$$

where $M(\xi)$ is analytic in the strip $-k < \mathrm{Im}(\xi) < k$. It is convenient to write (C.5) as

$$D_-(\xi)/C_+(\xi) = -Q(\xi)R(\xi), \tag{C.6}$$

where $R(\xi) = 2(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}$ and $Q(\xi) = M(\xi)/R(\xi)$. Now we can split $Q(\xi)$ and $R(\xi)$ into the product of factors analytic in the UHP and LHP, and as a result we obtain

$$\frac{D_-(\xi)}{Q_-(\xi)R_-(\xi)} = -C_+(\xi)Q_+(\xi) = P(\xi). \tag{C.7}$$

where $P(\xi)$ is an entire function. The factors $R_-(\xi)$, $R_+(\xi)$, and $Q_-(\xi)$ are given by

$$\begin{aligned} &R_-(\xi) = (\xi - ia_1)^{\frac{1}{2}} \qquad R_+(\xi) = 2(\xi + ia_2)^{\frac{1}{2}}, \\ &Q_-(\xi) = \exp\left\{-\frac{1}{2\pi i}\int_{-\infty}^{\infty} \frac{\ln Q(x)}{x - \xi}\,dx\right\} \qquad \text{for } \mathrm{Im}(\xi) < 0. \end{aligned} \tag{C.8}$$

Here $a_1 = \{(a^2 + 4\beta^2)^{\frac{1}{2}} - a\}/2$ and $a_2 = \{(a^2 + 4\beta^2)^{\frac{1}{2}} + a\}/2$. To determine $P(\xi)$ we must first anticipate that $Q_-(\xi) = O(1)$ as $\xi \to \infty$, and hence since $Q(\xi) \to 1$ as $\xi \to \infty$ the same is true of $Q_+(\xi)$. Also since $D_-(\xi)$ is proportional to the transform of the skin friction we expect $D_-(\xi) \sim 0(\xi)$ as $\xi \to \infty$. Using these order conditions plus Eqs. (C.8) and the information about $C_+(\xi)$ we see that $P(\xi)$ must be a constant, say P'. Then

$$C_+(\xi) = -\frac{P'}{Q_+(\xi)R_+(\xi)}, \qquad D_-(\xi) = P'Q_-(\xi)R_-(\xi). \tag{C.9}$$

To determine P' we substitute the above expression for $C_+(\xi)$ in Eq. (C.4). Noting that $R_+(\xi) \sim 2\xi^{\frac{1}{2}}$ and $Q_-(\xi) = 1$* plus terms of higher order for large ξ it is clear that $P' = a^{\frac{1}{2}}e^{-3\pi i/4}$. Consequently

$$\bar{\phi}_{yy}(\xi, 0) = e^{-3\pi i/4}a^{\frac{1}{2}}Q_-(\xi)R_-(\xi). \tag{C.10}$$

Appendix D. Evaluation of $I(\sigma)$. The functions $N_-(\xi)$ bnd $M_-(\xi)$ are defined by

$$N_-(\xi) \exp\left\{\frac{1}{2\pi i}\int_{\Gamma_2} \frac{\ln\,[(z^2 + \beta^2)^{\frac{1}{2}} + (z^2 + k^2)^{\frac{1}{2}}]}{z - \xi}\,dz\right\},$$

* As was pointed out earlier we anticipate that $Q_-(\xi)$ is 0(1) as $\xi \to \infty$; and since $Q_-(\xi)$ is determined only up to a constant multiplier we may take the leading term to be one. Actually the expression for $Q_-(\xi)$ as given in (C.8) does lead to the desired expression when ξ is large. This will be seen in Appendix E.

and

$$M_-(\xi) = \exp\left\{\frac{1}{2\pi i}\int_{\Gamma_2} \frac{\ln\left[(z^2 + iaz + \beta^2)^{\frac{1}{2}} + (z^2 + k^2)^{\frac{1}{2}}\right]}{z - \xi}\, dz\right\}.$$

If we restrict ξ so that $\mathrm{Im}(\xi) < 0$, then

$$\frac{N_-(\xi)}{M_-(\xi)} = \exp\left\{+\frac{1}{2\pi i}\int_{-\infty}^{\infty} \ln \frac{(x^2 + iax + \beta^2)^{\frac{1}{2}} + (x^2 + k^2)^{\frac{1}{2}}}{(x^2 + \beta^2)^{\frac{1}{2}} + (x^2 + k^2)^{\frac{1}{2}}} \frac{dx}{x - \xi}\right\}. \tag{D.1}$$

In the limit as $\xi \to 0$, our integration path must be indented to go above the origin; i.e. the path of integration is from $-\infty$ to $-\epsilon$, then a semi-circle of radius ϵ in the UHP, and then the real axis from ϵ to ∞. In the limit as $\epsilon \to 0$, the integration over the semi-circle is zero. Letting $r = ax/|\beta|^2$, $\sigma^2 = |\beta|^2/a^2 = \omega\nu/(cU_0)^2$, $N_-(0)/M_-(0) = \exp I(\sigma)$, and noting that as $k \to 0$, $(x^2 + k^2)^{\frac{1}{2}} = |x|$ we obtain from (D.1)

$$I(\sigma) = \frac{2\pi i}{1}\int_0^{\infty} \ln \frac{[i(1 + r) + \sigma^2 r^2]^{\frac{1}{2}} + \sigma r}{[i(1 - r) + \sigma^2 r^2]^{\frac{1}{2}} + \sigma r} \frac{dr}{r}. \tag{D.2}$$

This integral can be evaluated by first computing $dI/d\sigma$ and then integrating with respect to σ. We have

$$\begin{aligned} \frac{dI}{d\sigma} &= \frac{1}{2\pi i}\int_0^{\infty} \left\{\frac{1}{[\sigma^2 r^2 + ir + i]^{\frac{1}{2}}} - \frac{1}{[\sigma^2 r^2 - ir - i]^{\frac{1}{2}}}\right\} dr, \\ &= -\frac{1}{2\pi i\sigma} \ln \frac{\sigma + \gamma}{\sigma - \gamma}, \end{aligned} \tag{D.3}$$

where $\gamma = e^{\pi i/4}/2$. Consequently

$$I(\sigma) = \frac{1}{2\pi i}\int_{\sigma}^{\infty} \ln \frac{r + \gamma}{r - \gamma} \frac{dr}{r}. \tag{D.4}$$

The constant of integration (i.e. the upper limit in the above integral) is determined by noting that when $a = 0$, $N_-(0) = M_-(0)$ hence $I(\sigma) = 0$, but $a = 0$ corresponds to $\sigma = \infty$.

For $\sigma \geqq \frac{1}{2}$, $I(\sigma)$ can be evaluated directly by expanding the integrand in a power series in γ/σ and then integrating termwise. We obtain

$$I(\sigma) = -\frac{i}{\pi}\sum_{n=0}^{\infty} \frac{(\gamma/\sigma)^{2n+1}}{(2n+1)^2}, \qquad \sigma \geqq \frac{1}{2}. \tag{D.5}$$

For $\sigma < \frac{1}{2}$ the evaluation of (D.4) is slightly more difficult. It is convenient to write (D.4) as an integral from γ to ∞ plus an integral σ to γ, then we obtain

$$I(\sigma) = -\frac{1}{2}\ln(2\sigma) + \frac{1}{2\pi i}\int_{\sigma}^{\gamma} \frac{1}{r} \ln \frac{\gamma + r}{\gamma - r}\, dr. \tag{D.6}$$

After writing the integral in (D.6) as one from 0 to γ minus an integral from 0 to σ we obtain

$$I(\sigma) = -\frac{\pi i}{8} - \frac{1}{2}\ln(2\sigma) - \frac{1}{2\pi i}\int_0^r \frac{1}{r}\ln\frac{\gamma + r}{\gamma - r}\,dr$$
$$= -\frac{\pi i}{8} - \frac{1}{2}\ln(2\sigma) + \frac{i}{\pi}\sum_{n=0}^{\infty}\frac{(\sigma/\gamma)^{2n+1}}{(2n+1)^2}, \qquad \sigma \leqq \frac{1}{2}. \tag{D.7}$$

Appendix E. Asymptotic Expansion of $Q_-(\xi)$. The function $Q_-(\xi)$ as defined by (C.8) is

$$Q_-(\xi) = \exp\left\{-\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{\ln Q(x)}{x - \xi}\,dx\right\} \qquad \text{for } \mathrm{Im}(\xi) < 0, \tag{E.1}$$

where

$$Q(\xi) = \frac{(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}} + (\xi^2 + k^2)^{\frac{1}{2}}}{2(\xi^2 + ia\xi + \beta^2)^{\frac{1}{2}}}.$$

Except for branch points at $\xi = ik$ and $\xi = ia_1$, $Q(\xi)$ has no other singularities in the UHP. Hence using Cauchy's Integral Theorem the integral in (E.1) can be converted to an integral over T (see Fig. 6). So if we write $Q_-(\xi) = \exp\{-G(\xi)/2\pi i\}$ we have

$$G(\xi) = -\int_T \frac{\ln Q(z)}{z - \xi}\,dz. \tag{E.2}$$

For large ξ, $G(\xi)$ has asymptotic representation of the form

$$G(\xi) = G_0 + G_1/\xi + \cdots \tag{E.3}$$

where

$$G_0 = \lim_{\xi\to\infty} G(\xi), \qquad G_1 = -\lim_{\xi\to\infty}\xi^2\frac{dG}{d\xi}. \tag{E.4}$$

It is not difficult to show from (E.2) that G_0 is zero. Performing the limiting operation indicated in (E.4) we see that

$$G_1 = -\int_T \ln Q(z)\,dz. \tag{E.5}$$

It is convenient to write $Q(z)$ as

$$Q(z) = \frac{1}{2}\{1 + V(z); \qquad V(z) = \frac{(z + ik)^{\frac{1}{2}}(z - ik)^{\frac{1}{2}}}{(z - ia_1)^{\frac{1}{2}}(z + ia_2)^{\frac{1}{2}}}.$$

To evaluate (E.5) in the limit as $k \to 0$, we make the change of variables $x = z/ia_1$, and let $\delta = a_1/a_2$. Then our branch cut lies on the real axis from 0 to 1, and our path of integration is clockwise. The function $V(z)$ has the values

$$V(x) = \frac{xe^{-\pi i/2}}{(1 - x)^{\frac{1}{2}}(x + 1/\delta)^{\frac{1}{2}}}, \qquad V(x) = \frac{xe^{+\pi i/2}}{(1 - x)^{\frac{1}{2}}(x + 1/\delta)^{\frac{1}{2}}},$$

on the upper and lower sides of the branch cut. We obtain

$$G_1 = -ia_1 \int_0^{'} \ln \frac{1 + iq(x)}{1 - iq(x)}\, dx; \qquad q(x) = \frac{x}{(1-x)^{\frac{1}{2}}(x + 1/\delta)^{\frac{1}{2}}}. \tag{E.6}$$

The integral in Eq. (E.6) can be evaluated by integration by parts and some straightforward algebraic manipulation. We obtain

$$G_1 = -ia_1 \left\{ -\frac{\delta}{1-\delta}\pi i + 1 + \left(\frac{2}{1-\delta} \frac{1-\delta}{2i\delta^{\frac{1}{2}}} \ln \frac{1 + i\delta^{\frac{1}{2}}}{1 - i\delta^{\frac{1}{2}}} \right) \right\}.$$

So for ξ large

$$Q_-(\xi) = 1 - \frac{G_1}{2\pi i} \frac{1}{\xi} + O\left(\frac{1}{\xi^2}\right),$$

where G_1 is defined above.

REFERENCES

1. J. A. Lewis and G. F. Carrier, Some remarks on the flat plate boundary layer, Quart. App. Math., 7, pp. 228–233, (1949).
2. G. F. Carrier, On slow viscous flow, Report to the O.N.R. NR-062-163, March 1953.
3. G. F. Carrier, The mechanics of the Rijke tube, Quart. App. Math., *12*, pp. 383–395, (1955).
4. P. Lagestrom, J. Cole, and L. Trilling, Problems in the theory of viscous compressible fluids, GALCIT Report ONR Contract N6 ONR Task Order VIII, March 1949.
5. G. Doetsch, Theorie und anwendung der Laplace-transformation. Dover Publications, New York, N. Y., pp. 243–246 (1943).
6. H. Lamb, Hydrodynamics, 6th edition, Dover Publications, New York, N. Y., p. 620, (1945).
7. M. J. Lighthill, The response of laminar skin friction and heat transfer to fluctuations in the stream velocity, Proc. Roy. Soc. of London, A, *224* pp. 1–23, (1954).
8. F. K. Moore, Unsteady laminar boundary layer flow, NACA TN 2471 September, 1951.
9. S. Ostrach, Compressible laminar boundary layer and heat transfer for unsteady motions of a flat plate, NACA TN 3569, November, 1955.
10. J. T. Stuart, A solution of the Navier-Stokes and energy equations illustrating the response of skin friction and temperature of an infinite plate thermometer to fluctuations in the stream velocity, Proc. Roy. Soc. of London, A, *231*, pp. 116–130, (1955).
11. Lord Rayleigh, Theory of Sound, 2nd edition, Dover Publications, New York, N. Y., pp. 224–234, (1945).

Harvard University

(Received July 10, 1955; Revised April 13, 1956)

ONE-DIMENSIONAL SHOCK-MOTION OF A GAS WITH $\gamma = 3$, ENCLOSED BETWEEN FIXED WALLS*

By G. Kuerti and K. Faymon

1. Introduction. It is well known [1, 2, cf. also 3, p. 51 ff] that the continuous and homentropic one-dimensional motion of a perfect gas enclosed between fixed walls (motion in a "closed tube") results in infinite gradients of velocity and pressure at some point of the x, t-plane, whatever the initial conditions. The analysis of the motion beyond that breakdown point requires the introduction of a shock. Throughout the x, t-region after the shock, the solution is in general no longer homentropic. In the only exceptional case known, the regions of constant entropy on either side of the shock are also regions of *uniform* velocity and pressure.

In the example that follows here the region-wise homentropic character of the flow subsists, although the velocity is *non-uniform* in the two regions separated by the shock. This occurs in the motion within a "closed tube" of a perfect gas with $c_p/c_v = \gamma = 3$, if it is initially at rest and caused to move under an initially imposed linear density distribution. (With $\gamma = 3$ and constant entropy, this implies that the sound speed, c, the temperature, and the pressure, P, are initially linear, quadratic, and cubic functions of the station-coordinate along the tube, x.)

2. The continuous phase of the motion. We use dimensionless coordinates x and t, referring x to the length of the "tube", L, and t to L/C_0, where C_0 is the maximum initial sound speed realized at one end of the "tube". With u and c denoting the dimensionless particle and sound velocities referred to C_0, the quantities

$$r = \frac{1}{2}\left(u + \frac{2}{\gamma - 1}c\right), \qquad s = \frac{1}{2}\left(-u + \frac{2}{\gamma - 1}c\right) \tag{2.1}$$

are characteristic parameters for the general, one-dimensional, homentropic gas flow. The flow equations are linear in the independent variables r and s (cf. [4], p. 80):

$$x_r = (u - c)t_r, \qquad x_s = (u + c)t_s. \tag{2.2}$$

The initial and boundary conditions are

$$\begin{cases} c = 1 - \delta x, \quad u = 0 \quad \text{for} \quad 0 \leqq x \leqq 1 \quad \text{and} \quad t = 0; \\ u = 0 \quad \text{at} \quad x = 0 \quad \text{and} \quad x = 1 \quad \text{for all } t, \end{cases} \tag{2.3}$$

where $0 < \delta \leqq 1$, and $\gamma = 3$ in (2.1).

The boundary condition can be satisfied if the initial conditions for c and u

* This research was carried out with the support of the Office of Naval Research under Contract ONR Nonr-1141.

are extended to $-\infty < x < \infty$ by suitable reflections in the end faces. In the case under consideration we obtain instead of (2.3), for $t = 0$,

$$(2.4)\quad \begin{cases} c(x) = 1 - \delta x \quad \text{for} \quad 0 \leqq x \leqq 1, \qquad c(-x) = c(x) \quad \text{for} \quad -1 \leqq x \leqq 0, \\ c(x) = c(x \pm 2p) \quad \text{for all } x \text{ and integer } p; \\ u = 0 \quad \text{for all } x. \end{cases}$$

(The initial conditions are therefore only piecewise smooth).

For any $\gamma \neq 1$ and fairly general initial distributions of c and u, the solution of (2.2) can be found by an extension of Riemann's integration method due to Ludford [1]. Using that method, we have given [5] the complete solution of the problem $\{(2.2), (2.4); \gamma = 3\}$ with which we are here concerned; for our present purpose the following remarks will suffice.

For $\gamma = 3$, the characteristics C_+ and C_- in the x, t-plane are families of straight lines (cf. [1], p. 88). Since C_+ carries a constant value of r and C_- a constant value of s, and since $2r$ and $-2s$ are the slopes dx/dt of a C_+ and a C_-, respectively, the two characteristics through the field point x, t, with intercepts x_1 and x_2 ($x_1 < x_2$) on the x-axis, are given by

$$(2.5)\quad C_+ : (x - x_1)/t = 2r, \qquad C_- : (x - x_2)/t = -2s$$
$$\text{with } 2r = c(x_1) \text{ and } 2s = c(x_2),$$

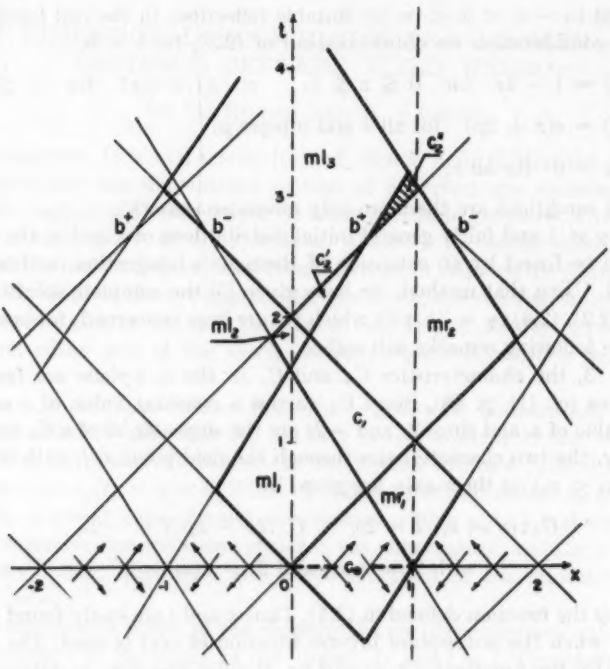
where $c(x)$ is the function defined in (2.4). Thus x and t are easily found in terms of r and s, when the *multivalued* inverse function of $c(x)$ is used. The analytic expressions for the functions $x(r, s)$ and $t(r, s)$ differ therefore in different quadrilateral regions of the x, t-plane bounded by pairs of those C_+ and C_- characteristics whose intercepts on the x-axis are pairs of consecutive integers (Fig. 1).

It can be easily seen that all C_-^+ characteristics emerging from a segment of the x-axis between integers p and $p + 1$ intersect at points ($x = p + \delta^{-1}$, $t = \pm\delta^{-1}$) for even p, and at points ($x = p + 1 - \delta^{-1}$, $t = \mp\delta^{-1}$) for odd p. Any such point marks the breakdown of the continuous phase of the motion (it represents here what in general is the point of smallest t-value on the envelope of one family of characteristics). Fig. 1 shows C_+ and C_- characteristics concurring in breakdown points b^+ and b^- respectively ($\delta = \frac{3}{8}$). The strip $0 \leqq x \leqq 1$ is covered by "central" quadrilaterals $c_0, c_1, \cdots$ with opposite corners on $x = 0$ and $x = 1$, and by left and right "marginal" quadrilaterals $ml_1, ml_2, \cdots$, $mr_1, mr_2, \cdots$, whose vertical diagonals coincide with $x = 0$ and $x = 1$, respectively.

Let n denote the largest integer smaller than δ^{-1},

$$(2.6)\qquad n = [1/\delta], \quad \text{or} \quad 1/\delta = n + \epsilon, \qquad 0 < \epsilon < 1,$$

where at present we exclude that δ^{-1} be an integer. If n is even (odd), the breakdown point is a $b^+(b^-)$ common to the quadrilaterals ml_{n+1}, c_n, $mr_n(ml_n, c_n, mr_{n+1})$. In either case, the quadrilateral c_n is "flipped over" at b so that the upper part of c_n (shaded area c_2'' in Fig. 1) is covered three times,



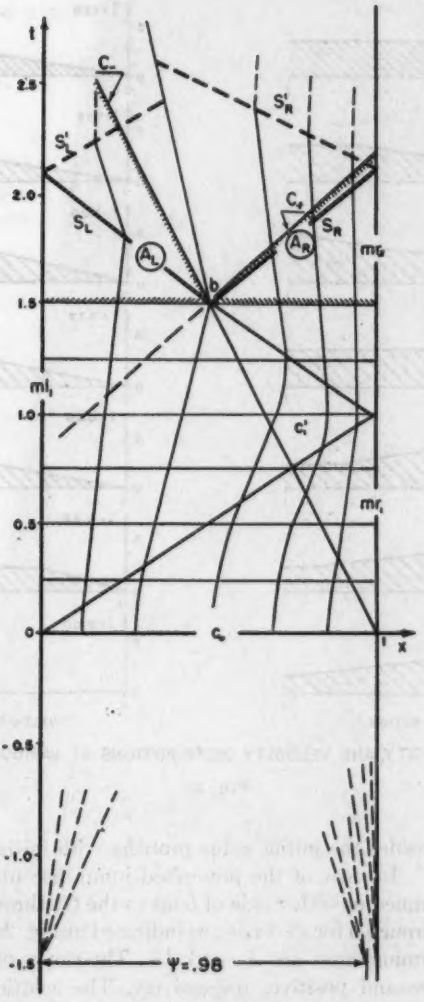
CHARACTERISTIC PATTERN FOR $\delta = 3/8$

FIG. 1

carrying three different solutions $x(r, s)$, $t(r, s)$. Up to $t = \delta^{-1}$, however, every point x, t carries only one solution.

In the marginal squares, t is always obtained as a function of $(r + s)$ or c. Thus c is a function of t alone, and since the particle acceleration $Du/Dt = -cc_x$, the particle lines must be straight. The particle lines in the central squares are obtained as arcs of conic sections. Fig. 2a shows the particle lines of this nonlinear oscillation for $\delta = \frac{3}{8}$. Within each of the central squares c_1, c_2, $\cdots$, c_{n-1} a complete reversal of the flow direction is effected, but the part of c_n below b (c_2' in Fig. 1) only reverses the flow of a part of the gas. The resulting c and u distribution at the instant of breakdown is thus of the type shown in Fig. 2b for $t = 1.50$. In general we have at $t = \delta^{-1}$, for *even* n

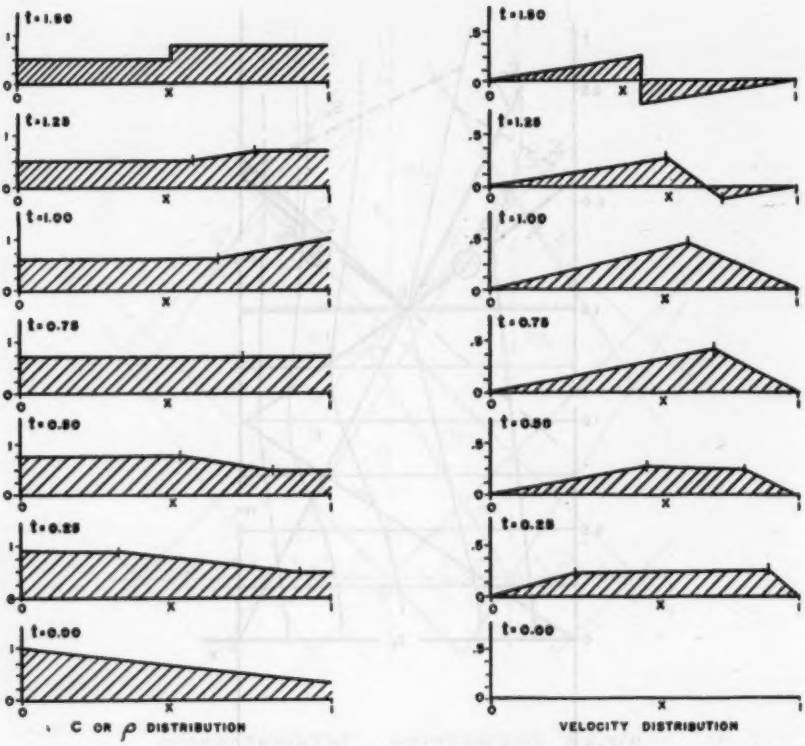
$$
\begin{array}{lll}
 & \text{left} & \text{right} \\
 & c = \tfrac{1}{2}[1 + \delta n] & c = \tfrac{1}{2}[1 + \delta(n-1)], \\
 & u = (\delta/2)x & u = (\delta/2)(x-1); \\
\end{array} \tag{2.7}
$$



x,t-DIAGRAM FOR δ=2/3

FIG. 2a

if n is *odd*, n and $n-1$ must be interchanged in (2.7). Thus the gas on the lefthand side of the breakdown cross section moves always to the right and vice versa, but c (together with density, temperature and pressure) is constant on either side of b and suffers at b a sudden decrease (increase) for even (odd) n.



SONIC SPEED, DENSITY, AND VELOCITY DISTRIBUTIONS AT VARIOUS TIMES FOR $\delta = 2/3$

FIG. 2b

We may now consider the initial value problem with initial conditions given by (2.7) for $t = \delta^{-1}$. In view of the prescribed jump discontinuity, continuous solutions are determined on either side of b up to the C_- through b for $x \to x_b-$ and up to the C_+ through b for $x \to x_b+$, as indicated in Fig. 2a. The corresponding regions of determinateness are A_L and A_R. The slopes of these C_- and C_+ are always negative and positive, respectively. The solution in A_L and A_R is of course identical with the solution of the original initial-value problem in ml_{n+1}, $mr_n(ml_n, mr_{n+1})$ for even (odd) n. For *even* n, it reads

$$\text{(2.8a)} \qquad x = \frac{(r-s)(1+\delta n)}{\delta(r+s)}, \qquad x = \frac{(r-s)[1+\delta(n-1)]}{\delta(r+s)} + 1,$$

$$\text{(2.8b)} \qquad t = \frac{1+\delta n}{\delta(r+s)} - \frac{1}{\delta}, \qquad t = \frac{1+(n-1)\delta}{\delta(r+s)} - \frac{1}{\delta}.$$

Solving for $c = r + s$ and $u = r - s$, we have

$$c = \frac{1 + \delta n}{1 + \delta t}, \qquad c = \frac{1 + \delta(n-1)}{1 + \delta t}, \tag{2.8c}$$

$$u = \frac{\delta x}{1 + \delta t}, \qquad u = \frac{\delta(x-1)}{1 + \delta t}. \tag{2.8d}$$

The formulas for odd n are again obtained by interchanging n and $n - 1$. The (straight) particle lines in $A_L(A_R)$ concur in $x = 0$ $(x = 1)$, $t = \delta^{-1}$ (Fig. 2a).

3. The shock solution. The pressure distribution in the "tube" at $t = \delta^{-1}$ is the same as that in the "head-on collision of two shocks." Thus it stands to reason that two shocks, S_L and S_R, should originate at b and be propagated into the regions A_L and A_R. The solutions behind S_L and S_R should abut each other in a discontinuity line through b, along which $u_L = u_R$ and $P_L = P_R$.

Let quantities behind the shock be marked by a circumflex. We then seek to determine the functions $\hat{x}(\hat{r}, \hat{s})$ and $\hat{t}(\hat{r}, \hat{s})$ in the form

$$\hat{x}(\hat{r}, \hat{s}) = \varkappa x(\hat{r}, \hat{s}) + \alpha, \tag{3.1a}$$

$$\hat{t}(\hat{r}, \hat{s}) = \varkappa t(\hat{r}, \hat{s}) + \beta, \tag{3.1b}$$

where x and t are the functions occurring in (2.8a, b). This form is suggested by the structure of Eqs. (2.2), which must be satisfied. Eq. (3.1) is meant to hold separately for the L-side and for the R-side. Six constants are therefore available at the outset.

Let U denote the shock velocity. If the shock separates regions of constant entropy, the shock Mach number $| U - u |/c$, which is a function of the entropy jump, must be constant along the shock, and so must the ratio $\hat{c}/c$. By (2.8b), t depends on c alone, hence $\hat{t}$ depends on $\hat{c}$ alone by (3.1). Expressing $\hat{c}/c$ in terms of t and $\hat{t}$ and setting $\hat{t} = t$ along the shock, we obtain a linear fractional function of t, which becomes independent of t, if

$$\delta\beta = \varkappa - 1. \tag{3.2}$$

When we introduce $(\varkappa - 1)/\delta$ for β in (3.1b) and use (2.8b) we obtain the function $\hat{c}(\hat{t})$ in the form

$$\hat{c} = \varkappa c(\hat{t}). \tag{3.3}$$

This clarifies the physical meaning of $\varkappa$. Again, (3.3) holds in A_R and A_L, possibly with different values of $\varkappa$.

We now introduce (3.3) in (3.1a), solve for $\hat{u}$ and obtain (e.g. for even n)

$$\hat{u}_L = \frac{\delta}{1 + \delta n} c_L(\hat{t})[\hat{x} - \alpha_L], \qquad \hat{u}_R = \frac{\delta}{1 + \delta(n-1)} c_R(\hat{t})[\hat{x} - (\alpha_R + \varkappa_R)]. \tag{3.4}$$

Along the discontinuity line $\hat{u}_L = \hat{u}_R$. By means of (2.8c) we so obtain an identity in $\hat{x}$:

$$\hat{x} - \alpha_L = \hat{x} - (\alpha_R + \kappa_R) \quad \text{or} \quad \alpha_L = \kappa_R + \alpha_R \equiv \psi, \tag{3.5}$$

where ψ denotes the common value of α_L and $\alpha_R + \kappa_R$. Introduction of (3.5) in (3.4) yields the same expression in both regions behind S_L and S_R, viz.

$$\hat{u} = \hat{u}_L = \hat{u}_R = \frac{\delta(\hat{x} - \psi)}{1 + \delta\hat{t}}. \tag{3.6}$$

Relations (3.5) and (3.6) hold for even and odd n.

The constant ψ is determined by imposing the condition of equal pressure along the discontinuity line. We first note that for $\gamma = 3$

$$U = \hat{u} \pm \sqrt{(\hat{u} - u)^2 + c^2}, \tag{3.7}*$$

where the plus sign holds for U_R and the minus sign for U_L. From (3.7), (3.6) with $\hat{x} = x$ and $\hat{t} = t$, (2.8c, d) and $M_L = (u_L - U_L)/c_L$, $M_R = -(u_R - U_R)/c_R$, we obtain M_L and M_R in terms of the constants δ, n and ψ, *independent of x and t.* These results are now introduced in the expression for the pressure ratio across the shock, which for $\gamma = 3$ is given by

$$\hat{P}_{L,R}/P_{L,R} = \tfrac{1}{2}(3M_{L,R}^2 - 1). \tag{3.8}$$

Here the pressures P_L and P_R ahead of the shock are

$$P_{L,R} = [C_0 c_{L,R}(t)]^3/\sqrt{27}\mathfrak{C} \tag{3.9}$$

where $\mathfrak{C} = P_0/\rho_0^3$ is associated with the entropy value in the continuous phase of the motion, and P_0, ρ_0, C_0 are the values of these parameters at $x = t = 0$. Thus we find the pressure *immediately behind the shock at the time t*

$$\left.\begin{matrix}\hat{P}_L^* \\ \hat{P}_R^*\end{matrix}\right\} = \frac{[C_0 c_{L,R}]^3}{\sqrt{12}\mathfrak{C}} \times \begin{matrix}F(\psi, n) \\ F(1 - \psi, n - 1)\end{matrix}, \tag{3.10}$$

$$\text{with } F(\psi, n) = \left[\frac{\delta\psi + \sqrt{(\delta\psi)^2 + (1 + \delta n)^2}}{1 + \delta n}\right]^2 - \frac{1}{3}.$$

Now $\hat{P}_{L,R}$ is a function of $\hat{c}_{L,R}$ in homentropic flow and $\hat{c}$ depends only on $\hat{t}$ by (3.3) and (2.8c). Thus the functions $\hat{P}_{L,R}$ depend only on $\hat{t}$ and are therefore identical with the expressions (3.10) if t is replaced by $\hat{t}$. Identification of $\hat{P}_L$ and $\hat{P}_R$ along the discontinuity line (i.e. for the same $\hat{t}$) gives finally for even n

$$[1 + \delta n]^2 F(\psi, n) = [1 + \delta(n - 1)]^2 F(1 - \psi, n - 1), \tag{3.11}$$

n and $n - 1$ to be interchanged, if n is odd. A simple discussion shows that this equation has only one root, ψ^*; for even n, $0 < \psi_e^* < \frac{1}{2}$ and for odd n, $\frac{1}{2} < \psi_0^* < 1$; also $\psi^* \to \frac{1}{2}$ with increasing n.

With $\hat{u}$ now determined by (3.6), the particle lines behind the shock become $x - \psi^* = a(t + \delta^{-1})$, see also Fig. 2a. On the discontinuity line, $2a = 1 - \delta[n + \psi_e^*]$ if n is even and $2a = -1 + \delta[(n + 1) - \psi_0^*]$ if n is odd. Likewise M_L and M_R may be computed. Finally $\hat{u}$, $u_{L,R}$ and $c_{L,R}$ in (3.7) are expressed

* For (3.7) and (3.8) see e.g. [4], p. 149.

in terms of $\hat{x}$, $\hat{t}$ and x, t; identification of $\hat{x}$ with x and $\hat{t}$ with t and replacement of $U_{L,R}$ by dx/dt yields the differential equations of the shocks. Both shocks are straight.

The quantities κ_L and κ_R can be found from the formula that expresses $\hat{c}/c$ in terms of M with the result

$$\kappa_L^2 = \tfrac{1}{3} + F(\psi_0^*, n) - [\delta\psi_0^*/(1 + \delta n)]^2$$
$$\kappa_R^2 = \tfrac{1}{3} + F(1 - \psi_0^*, n - 1) - [\delta(1 - \psi_0^*)/(1 + \delta(n - 1))]^2 \tag{3.12}$$

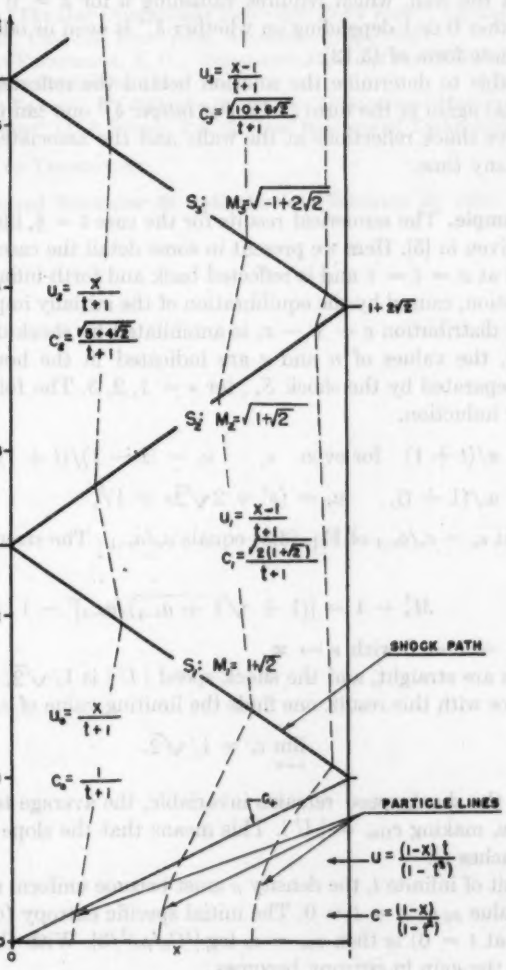


FIG. 3

for even n. The values for odd n are obtained by changing ψ_*^* into ψ_0^*, and n into $n - 1$ in κ_L and into $n + 1$ in κ_R.

Since only two shock conditions (viz. the equations relating $\hat{P}/P$ and $\hat{c}/c$ to M) have been employed explicitly, it must be checked that a third shock condition is satisfied. This has been done [5, p. 26].

When δ^{-1} is an integer, the breakdown point occurs at $x = 0(1)$ for even (odd)δ^{-1}. In this case only one shock will enter the region $0 < x < 1$, the other one passes symmetrically into the region $-1 < x < 0$ $(1 < x < 2)$, and there is no discontinuity line. The value of ψ^* is then determined by the boundary condition at the wall, which requires vanishing u for $x = 0$ (or $x = 1$). This makes ψ^* either 0 or 1 depending on whether δ^{-1} is even or odd; κ is found from the appropriate form of (3.12).

It is possible to determine the solution behind the reflected shocks (S_L' and S_R' in Fig. 2a) again in the form (3.1). For *integer* δ^{-1} one can therefore calculate the successive shock reflections at the walls and the associated u- and c-distributions for any time.

4. An example. The numerical results for the case $\delta = \frac{1}{3}$, illustrated in Fig. 2, have been given in [5]. Here we present in some detail the case $\delta = 1$, where the shock starts at $x = t = 1$ and is reflected back and forth infinitely many times, until the motion, caused by the equilibration of the initially imposed sound speed (or density) distribution $c = 1 - x$, is annihilated by shock-damping.

In Fig. 3, the values of u and c are indicated in the homentropic regions $(\nu - 1)$, ν separated by the shock S_ν, for $\nu = 1, 2, 3$. The following results are obtained by induction.

$$(4.1)\quad u_\nu = x/(t+1) \quad \text{for even} \quad \nu, \qquad u_\nu = (x-1)/(t+1) \quad \text{for odd} \quad \nu$$

$$(4.2)\quad c_\nu = a_\nu/(1+t), \qquad a_\nu = (\nu^2 + 2\sqrt{2}\nu + 1)^{\frac{1}{2}}, \qquad \nu = 0, 1, 2, \cdots$$

The constant $\kappa_\nu = c_\nu/c_{\nu-1}$ of Eq. (3.3) equals $a_\nu/a_{\nu-1}$. The strength of the shocks is given by

$$(4.3)\qquad M_\nu^2 - 1 = [(1 + \sqrt{1 + a_{\nu-1}})/a_{\nu-1}]^2 - 1$$

Clearly, $M_\nu^2 - 1 \to 0$ with $\nu \to \infty$.

All shocks are straight, and the shock speed $|U|$ is $1/\sqrt{2}$, independent of ν. In accordance with this result, one finds the limiting value of c_ν at $t = 1 + \nu\sqrt{2}$

$$(4.4)\qquad \lim_{\nu\to\infty} c_\nu = 1/\sqrt{2}.$$

Thus, while the shock speed remains invariable, the average temperature of the gas increases, making $c_{\text{lim}} = |U|$. This means that the slope of the characteristics approaches U.

In the limit of infinite t, the density ρ must become uniform and hence assume half of its value ρ_0 at $x = t = 0$. The initial specific entropy (constant throughout the gas at $t = 0$) is then $s_{\text{in}} = c_v \log [(C_0/\rho_0)^2/3]$. With $C_{\text{fin}} = C_0/\sqrt{2}$ and $\rho_{\text{fin}} = \rho_0/2$, the gain in entropy becomes

(4.5) $$s_{\text{fin}} - s_{\text{in}} = c_v \log 2,$$

the same as in a Gay-Lussac process resulting in half of the (initially uniform) density.

REFERENCES

1. Ludford, Geoffrey S. S., "On an Extension of Riemann's Method of Integration with Application to One-Dimensional Gas Dynamics", Proc. Cambridge Philos. Soc., vol. 48, pp. 499–510 (1952).
2. Schot, Steven H., "Isentropic Motion of a Monatomic Gas in a Closed Tube", OOR Report T.R. 2, University of Maryland, Oct. 1953.
3. Zaldastani, O., "The One-Dimensional Isentropic Fluid Flow", Advances in Applied Mechanics, Vol. III, New York, Academic Press, 1953.
4. Courant, R. and Friedrichs, K. O., "Supersonic Flow and Shock Waves", New York, Interscience Publishers, 1948.
5. Kuerti, G., and Faymon, K., "Shocks in the One Dimensional Motion of a Gas in a Closed Tube", ONR Report T.R. 2, Case Institute of Technology, April, 1955.

Case Institute of Technology

(Received November 29, 1955; Revised February 26, 1956)

OSCULATORY INTERPOLATION AND INTEGRATION

By J. C. Cooke

Introdution. Formulas are given for interpolation and integration of a function $f(x)$, which is tabulated, together with its first, or its first two, or its first three derivatives, at equal intervals of the argument x. Such tabulations occur quite often in the numerical integration of a differential equation. Especially powerful formulas for integration and for the special case of half-way interpolation may be derived if derivatives up to the second are known. Illustrations are given from a boundary layer equation.

Notation. We shall use the notation of Hartree (3) and will obtain the results by the free use of operators. Central differences will be denoted by powers of δ, forward differences by Δ and backward differences by ∇. The function concerned has values $f_0, f_1, f_2, \cdots$, tabulated at equal intervals h of the argument $x_0, x_1, x_2, \cdots$, that is $f(x_r) = f_r$. Other operators are defined by

$$Ef_r = f_{r+1}, \qquad Uf_r = hf_r',$$

$$\mu\delta^n f_r = \tfrac{1}{2}(\delta^n f_{r-\frac{1}{2}} + \delta^n f_{r+\frac{1}{2}}),$$

primes denoting differentiation with respect to x. Other formulas that we shall use are

$$\begin{aligned} &1 + \Delta = E = e^U, \qquad \mu^2 = 1 + \tfrac{1}{4}\delta^2, \\ &\delta = E^{\frac{1}{2}} - E^{-\frac{1}{2}}, \qquad 2\mu = E^{\frac{1}{2}} + E^{-\frac{1}{2}}, \\ &\delta^n f_0 + \delta^n f_1 = 2(\mu\delta^{n-1} f_1 - \mu\delta^{n-1} f_0). \end{aligned} \tag{1}$$

These are given by Hartree.

Operational expansions which will be required are

$$\delta/\mu U = 1 - \frac{1}{12}\delta^2 + \frac{11}{720}\delta^4 - \frac{191}{60480}\delta^6 + \frac{2497}{3628800}\delta^8 + \cdots,$$

$$\mu\delta/U = 1 + \frac{1}{6}\delta^2 - \frac{1}{180}\delta^4 + \frac{1}{1512}\delta^6 - \frac{23}{226800}\delta^8 + \cdots,$$

$$\delta^2/U^2 = 1 + \frac{1}{12}\delta^2 - \frac{1}{240}\delta^4 + \frac{31}{60480}\delta^6 - \frac{289}{3628800}\delta^8 + \cdots,$$

$$\delta^3/\mu U^3 = 1 + \frac{1}{240}\delta^4 - \frac{31}{30240}\delta^6 + \frac{289}{1209600}\delta^8 \cdots.$$

These are obtained from Bickley (1), who gives general formulas for $(U/\delta)^n$ and $(U/\delta)^n/\mu$.

In addition we shall require the expansion, given by Hartree

$$\sinh \theta U/\sinh U = \theta + E_1^{\mathrm{II}}(\theta)\delta^2 + E_1^{\mathrm{IV}}(\theta)\delta^4 + \cdots,$$

where the E's are the Everett coefficients in the interpolation formula which bears his name.

The values of the E's are

$$E_1^{II}(\theta) = \theta(\theta^2 - 1^2)/3!, \qquad E_1^{IV}(\theta) = \theta(\theta^2 - 1^2)(\theta^2 - 2^2)/5!,$$
$$E_1^{VI}(\theta) = \theta(\theta^2 - 1^2)(\theta^2 - 2^2)(\theta^2 - 3^2)/7!,$$

and so on. If we write $\bar{\theta} = 1 - \theta$, $E_0^{(r)}(\theta)$ is the same function of $\bar{\theta}$ as $E_1^{(r)}(\theta)$ is of θ.

The aim will be to push the formulas as far as possible towards the higher derivatives, as this will lead to higher accuracy. We shall not take differences, beyond the first, of the function itself or of its intermediate derivatives. There is some choice open to us here, and we shall attempt to make the coefficients of the differences of the highest derivative as small as possible; in fact one coefficient can be made equal to zero if we so desire.

We proceed now to the various cases. The first will be dealt with in some detail; in later ones the proofs will be more sketchy, but the details may easily be filled in.

Interpolation given the first derivative. We seek to find operators $\phi_0(\delta)$ and $\phi_1(\delta)$ such that

$$f(x + \theta h) = (1 - \theta)f_0 + \theta f_1 + h\mu\delta\phi_0(\delta)f_0' + h\mu\delta\phi_1(\delta)f_0',$$

the last two terms being a correction to linear interpolation.

In operational form we have

$$e^{\theta U} = 1 - \theta + \theta e^{U} + \mu\phi_0(\delta)\delta U + \mu\phi_1(\delta)\delta e^{U}U.$$

The operators $\phi(\delta)$ will be even functions of δ as we shall see. μ is also an even function of δ. Changing the sign of δ and therefore of U gives

$$e^{-\theta U} = 1 - \theta + \theta e^{-U} + \mu\phi_0(\delta)\delta U + \mu\phi_1(\delta)\delta e^{-U}U,$$

and consequently we have on subtraction

$$\phi_1(\delta) = (\sinh \theta U - \theta \sinh U)/\mu\delta U \sinh U$$
$$= \frac{1}{\delta^2}\frac{\delta}{\mu U}\left\{\frac{\sinh \theta U}{\sinh U} - \theta\right\}.$$

Hence we have on expanding and multiplying out

$$f(x_0 + \theta h) = (1 - \theta)f_0 + \theta f_1 + h(E_0^{II}\mu\delta f_0' + E_1^{II}\mu\delta f_1' + F_0^{III}\mu\delta^3 f_0' + F_1^{III}\mu\delta^3 f_1' + F_0^{V}\mu\delta^5 f_0' + F_1^{V}\mu\delta^5 f_1') + 0(h^5), \tag{2}$$

where

$$F_0^{III} = E_0^{IV} - \frac{1}{12}E_0^{II} = \bar{\theta}(\bar{\theta}^2 - 1)(3\bar{\theta}^2 - 17)/360,$$

$$F_0^{V} = E_0^{VI} - \frac{1}{12}E_0^{IV} + \frac{11}{720}E_0^{II} = \bar{\theta}(\bar{\theta}^2 - 1)(6\bar{\theta}^4 - 99\bar{\theta}^2 + 377)/30240,$$

together with similar formulas for F_1^{III} and F_1^{V} with $\bar{\theta}$ replaced by θ.

Interpolation given the first two derivatives. Writing the operator

$$e^{\theta U} = 1 - \theta + \theta e^{U} + \phi_0(\delta)U^2 + \phi_1(\delta)U^2e^{U}$$

leads in a similar manner as before to

$$\phi_1(\delta) = \frac{1}{\delta^2}\frac{\delta^2}{U^2}\left\{\frac{\sinh \theta U}{\sinh U} - \theta\right\}.$$

Hence we have

$$\begin{aligned} f(x + \theta h) = (1 - \theta)f_0 + \theta f_1 + h^2(E_0^{II}f_0'' + E_1^{II}f_1'' \\ + G_0^{II}\delta^2 f_0'' + G_1^{II}\delta^2 f_1'' + G_0^{IV}f_0'' + G_1^{IV}f_1'') + 0(h^8), \end{aligned} \tag{3}$$

where

$$G_0^{II} = E_0^{IV} + \frac{1}{12}E_0^{II} = \bar{\theta}(\bar{\theta}^2 - 1)(3\bar{\theta}^2 - 7)/360,$$

$$G_0^{IV} = E_0^{VI} + \frac{1}{12}E_0^{IV} - \frac{1}{240}E_0^{II} = \bar{\theta}(\bar{\theta}^2 - 1)(2\bar{\theta}^4 - 19\bar{\theta}^2 + 37)/10080.$$

We could add and subtract a term $a_0\mu\delta f_0' + a_1\mu\delta f_1'$ to this, expressing the one subtracted in terms of differences of f_0 and f_1 and then attempt to choose a_0 and a_1 so that the resulting differences have small coefficients. It does not, however, seem very obvious how to choose the a's and we shall not attempt it here. In the case of half-way interpolation and integration we shall find this device very powerful.

Interpolation given the first three derivatives. The same methods lead to the result

$$\begin{aligned} f(x_0) + \theta h) = (1 - \theta)f_0 + \theta f_1 + h^2(E_0^{II}f_0'' + E_1^{II}f_1'') \\ + h^3(H_0^{I}\mu\delta f_0''' + H_1^{I}\mu\delta f_1''' + H_0^{III}\mu\delta^3 f_0''' + H_1^{III}\mu\delta^3 f_1''') + 0(h^8), \end{aligned} \tag{4}$$

where

$$H_0^{I} = E_0^{IV} + \frac{1}{12}E_0^{II} = \bar{\theta}(\bar{\theta}^2 - 1)(3\bar{\theta}^2 - 7)/360,$$

$$H_0^{III} = E_0^{VI} - \frac{1}{90}E_0^{II} = \bar{\theta}(\bar{\theta}^2 - 1)(3\bar{\theta}^4 - 39\bar{\theta}^2 + 80)/15120.$$

Half-way Interpolation. In the case $\theta = \bar{\theta} = \frac{1}{2}$ we have

$$E^{II} = -\frac{1}{16}, \quad E^{IV} = \frac{3}{256}, \quad E^{VI} = -\frac{5}{2048}, \quad E^{VIII} = \frac{35}{32768}.$$

We note also that

$$\mu\delta f_0' + \mu\delta f_1' = 2\mu^2\delta f_{\frac{1}{2}}' = (2\delta + \tfrac{1}{24}\delta^3)f_{\frac{1}{2}}',$$

with a similar result for higher orders.

Hence we obtain from equations (2), (3) and (4)

$$f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1) + h\left(-\frac{1}{8}\delta f'_{\frac{1}{2}} + \frac{1}{384}\delta^3 f'_{\frac{1}{2}} - \frac{13}{30080}\delta^5 f'_{\frac{1}{2}}\right) + 0(h^8), \tag{5}$$

$$f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1) + h^2\left\{-\frac{1}{16}(f_0'' + f_1'') + \frac{5}{768}(\delta^2 f_0'' + \delta^2 f_1'') - \frac{37}{30720}(\delta^4 f_0'' + \delta^4 f_1'')\right\} + 0(h^8), \tag{6}$$

$$f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1) - \frac{1}{16}h^2(f_0'' + f_1'') + h^3\left(\frac{5}{384}\delta f_{\frac{1}{2}}''' - \frac{11}{46080}\delta^3 f_{\frac{1}{2}}'''\right) + 0(h^8). \tag{7}$$

The device mentioned above enables us to improve equations (6) and (7) considerably.

In equation (6) we add and subtract $ah\delta f'_{\frac{1}{2}}$. We have

$$ah\delta f'_{\frac{1}{2}} = ah\mu\left(\frac{\delta}{\mu U}\right)Uf'_{\frac{1}{2}} = ah^2\mu\left(\frac{\delta}{\mu U}\right)f''_{\frac{1}{2}}$$

$$= ah^2\mu\left(1 - \frac{1}{12}\delta^2 + \frac{11}{720}\delta^4 - \cdots\right)f''_{\frac{1}{2}}.$$

Hence from equation (6) we have

$$f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1) - ah\delta f'_{\frac{1}{2}} + ah^2\mu\left(1 - \frac{1}{12}\delta^2 + \frac{11}{720}\delta^4 - \cdots\right)f''_{\frac{1}{2}} + h^2\mu\left(-\frac{1}{8} + \frac{5}{384}\delta^2 - \frac{37}{15360}\delta^4 + \cdots\right)f''_{\frac{1}{2}}.$$

We now choose a so that the coefficient of $\mu\delta^2 f''_{\frac{1}{2}}$ vanishes. This gives $a = 5/32$. This value considerably reduces the magnitude of coefficient of $h^2\mu\delta^4 f''_{\frac{1}{2}}$, in fact it is changed from $-37/15360$ to $-1/46080$. It can also be shown that the coefficient $\mu\delta^6 f''_0$ is reduced by a factor of about 80.

Hence we have, using the value of a given above,

$$f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1) - \frac{5}{32}h\delta f'_{\frac{1}{2}} + \frac{1}{64}h^2(f_0'' + f_1'') - \frac{1}{92160}h^3(\delta^4 f_0'' + \delta^4 f_1'') + 0(h^8), \tag{8}$$

which is a very powerful formula, normally only requiring the first three terms.

From this we obtain at once

$$f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1) - \frac{5}{32}h\delta f'_{\frac{1}{2}} + \frac{1}{64}h^2(f_0'' + f_1'') - \frac{1}{46080}h^3\delta^3 f_0''' + 0(h^8). \tag{9}$$

Integration. We note that

$$\int_{x_0}^{x_1} f'(x)\,dx = f_1 - f_0 = \delta f_{\frac{1}{2}}.$$

Hartree gives the result

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{1}{2}h(f_0+f_1) + h^2\left\{-\frac{1}{12}(f_1'-f_0') + \frac{1}{720}(\delta^2 f_1' - \delta^2 f_0') - \frac{1}{6720}(\delta^4 f_1' - \delta^4 f_0')\right\} + 0(h^9) \tag{10}$$

which is the required form when the first derivative is given.

If the first two derivatives are given we write

$$\delta f_{\frac{1}{2}} = h\mu f_{\frac{1}{2}}' - ah^2\delta f_{\frac{1}{2}}'' + h^3\phi(\delta)\mu f_{\frac{1}{2}}''',$$

so that

$$\phi(\delta) = \frac{1}{\delta^2}\left(\frac{\delta^3}{\mu U^3} - \frac{\delta^2}{U^2}\right) + a\frac{\delta}{\mu U}$$
$$= -\frac{1}{12} + \frac{1}{120}\delta^2 - \frac{31}{20160}\delta^4 + \cdots + a\left(1 - \frac{1}{12}\delta^2 + \frac{11}{720}\delta^4 - \cdots\right).$$

We choose a so as to make the coefficient of δ^2 vanish. Hence we take $a = \frac{1}{12}$. On lowering the order of differentiation by one we have

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{1}{2}h(f_0+f_1) - \frac{1}{10}h^2\delta f_{\frac{1}{2}}' + h^3\left(\frac{1}{60}\mu f_{\frac{1}{2}}'' - \frac{1}{100800}\mu\delta^4 f_{\frac{1}{2}}''\right) + 0(h^9). \tag{11}$$

When the integration is over a number of intervals it is better to write this in the form

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{1}{2}h(f_0+f_1) - \frac{1}{10}h^2(f_1'-f_0') + h^3\left\{\frac{1}{120}(f_0''+f_1'') - \frac{1}{100800}(\mu\delta^2 f_1'' - \mu\delta^2 f_0'')\right\} + 0(h^9). \tag{12}$$

This is not ideal for numerical work, owing to the presence of the third term on the right hand side, but it is so powerful that it may be worth the labour involved.

This term can of course be removed by choosing $a = \frac{1}{12}$. We then obtain

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{1}{2}h(f_0+f_1) - \frac{1}{12}h^2(f_1'-f_0') + h^3\left\{\frac{1}{720}(\mu\delta f_1'' - \mu\delta f_0'') - \frac{1}{3780}(\mu\delta^3 f_1'' - \mu\delta^3 f_0'')\right\} + 0(h^9). \tag{13}$$

Finally when the first three derivatives are given we get

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{1}{2}h(f_0+f_1) - \frac{1}{10}h^2(f_1'-f_0') + \frac{1}{120}h^3(f_0''+f_1'') - \frac{1}{100800}h^4(\delta^2 f_1''' - \delta^2 f_0''') + 0(h^9), \tag{14}$$

or alternatively

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{1}{2}h(f_0 + f_1) - \frac{1}{12}h^2(f_1' - f_0') + h^4\left\{\frac{1}{720}(f_1''' - f_0''') - \frac{1}{30240}(\delta^2 f_1''' - \delta^2 f_0''')\right\} + 0(h^6), \tag{15}$$

which may be more suitable for a number of intervals.

Illustrations. We shall illustrate our results by consideration of the Blasius equation

$$f''' + ff'' = 0$$

with the boundary conditions $f(0) = f'(0) = 0, f'(\infty) = 2$, taking the solution given by Cope and Hartree (**2**). We shall tabulate this at intervals 0.4 of the argument η, (Table I) and attempt to subtabulate in the range 0.8 to 1.2.

We first find the values for $\eta = 1.0$, and then again by two more subtabulations those for $\eta = 0.9$ and $\eta = 1.1$.

The results are given in Table II. The value of $f'''(1)$ is not good, but it is sufficient for the next subtabulation. It will be seen that the values improve the more derivatives available. A better value for $f'''(1)$ could have been obtained from the differential equation but this was deliberately not done. The values of $f''(0.9)$ and $f''(1.1)$ could have been improved if subtabulation had proceeded above and below, so as to have had available the third differences of f'''.

TABLE I

η	f	f'	f''	f'''
0.0	0.000000	0.000000	1.328232	0.000000
0.4	0.106108	0.529420	1.309553	−0.138954
0.8	0.420320	1.033519	1.186649	−0.498772
1.2	0.922289	1.457971	0.912371	−0.841470
1.6	1.569091	1.752168	0.556521	−0.873232
2.0	2.305740	1.911036	0.256937	−0.592430

TABLE II

η	f	E	f'	E
0.9	0.529517	0	1.149522	0
1.0	0.650024	−1	1.259537	+1
1.1	0.781193	−1	1.362627	+1
η	f''	E	f'''	E
0.9	1.131686	+35		
1.0	1.066992	+13	−0.693394	−184
1.1	0.992384	+22		

The errors E are given as figures in the sixth decimal place.

To check the integration formula we evaluate

$$I = \int_0^2 f'(\eta)\, d\eta$$

using equation (12) ignoring the last term. We obtain the result $I = 2.305754$, instead of 2.305740, an error of -14 in the sixth decimal place.

REFERENCES

(1) Bickley, W. G. Difference and associated operators, with some applications. *Journ. of Math. and Phys.* **27** (1948), 183.

(2) Cope, W. F. and Hartree, D. R. The Laminar Boundary Layer in Compressible Flow. *Phil. Trans. Roy. Soc.* A. **241** (1948), 64.

(3) Hartree, D. R. *Numerical Analysis*, Oxford, 1952.

University of Malaya

(Received January 17, 1956)

LIMITS OF THE CHARACTERISTIC VALUES FOR CERTAIN BOUNDARY-VALUE PROBLEMS ASSOCIATED WITH DIFFERENCE EQUATIONS

By Tomlinson Fort

We shall consider the difference equation

$$\sum_{v=1}^{\mu} \Delta_{i_v}\{k_v(i_1, i_2, \cdots, i_{v-1}, i_v - 1, i_{v+1}, \cdots, i_\mu)\cdot$$

$$\Delta_{i_v} y(i_1, i_2, \cdots, i_{v-1}, i_v - 1, i_{v+1}, \cdots, i_\mu)\} \tag{1}$$

$$+ [\lambda n^{-2} - s(i_1, i_2, \cdots, i_v, \cdots i_\mu)]y(i_1, i_2, \cdots, i_v, \cdots, i_\mu) = 0$$

We assume that $i_1, i_2, \cdots, i_\mu$ are limited to integral values and that k and s are defined when

$$1 \leqq i_q \leqq m_q, \qquad q = 1, 2, \cdots, \mu \tag{2}$$

and that

$$k > 0, \qquad s > 0$$

We also assume that

$$m_q = M_q \cdot n \tag{3}$$

where M_q is independent of n.

We let

$$m = m_1 m_2 \cdots m_\mu, \qquad M = M_1 M_2 \cdots M_\mu \tag{4}$$

We assume in addition that

$$k_v(i_1, i_2, \cdots, i_\mu) = \bar{k}_v(\chi_i^{(1)}, \chi_i^{(2)}, \cdots, \chi_i^{(\mu)})$$

where

$$\chi_{i+1}^{(q)} - \chi_i^{(q)} = 1/n, \qquad q = 1, 2, \cdots, \mu, \; i = 1, 2, \cdots, m_q \tag{5}$$

We assume moreover that

$$0 < \lim_{n\to\infty} \bar{k}_v(\chi_i^{(1)}, \chi_i^{(2)}, \cdots, \chi_i^{(\mu)}) = K_v(\chi_1, \chi_2, \cdots, \chi_\mu) < L$$

where L is independent of v and n and where K_v is defined and continuous

$$0 \leqq \chi_q \leqq M_q, \qquad q = 1, 2, \cdots, \mu. \tag{6}$$

Like assumptions are made about s and a continuous function S.

In case we consider the differential equation analagous to (1) we naturally assume the existence of the first partial derivatives of K_v.

Interesting special cases of (1) are

$$\Delta\{k(i-1)\Delta y(i-1)\} + (\lambda n^{-2} - s(i))y(i) = 0$$

which is the equation determining the shape of a loaded vibrating string[1]; and

$$\Delta_i\{b(i-1,j)\Delta_i y(i-1,j)\} + \Delta_j\{k(i,j-1)\Delta_j y(i,j-1)\} + [\lambda n^{-2} - s(i,j)]y(i,j) = 0$$

which is the equation determining the shape of a two dimensional loaded vibrating net.[2]

Equation (1) is written in difference form. We now write it in ordinary form. This is as follows:

$$\begin{aligned}
&\textstyle\sum_{\nu=1}^{\mu} k_\nu(i_1, i_2, \cdots, i_{\nu-1}, i_\nu - 1, i_{\nu+1}, \cdots, i_\mu)\cdot \\
&\quad y(i_1, i_2, \cdots, i_{\nu-1}, i_\nu - 1, i_{\nu+1}, \cdots, i_\mu) \\
&\quad + \textstyle\sum_{\nu=1}^{\mu} k_\nu(i_1, i_2, \cdots, i_{\nu-1}, i_\nu, i_{\nu+1}, \cdots, i_\mu)\cdot \\
&\quad\quad y(i_1, i_2, \cdots, i_{\nu-1}, i_\nu + 1, i_{\nu+1}, \cdots, i_\mu) \\
&\quad\quad + [\lambda n^{-2} - s(i_1, i_2, \cdots, i_\mu) \\
&\quad - \textstyle\sum_{\nu=1}^{\mu} \{k_\nu(i_1, i_2, \cdots, i_{\nu-1}, i_\nu - 1, i_{\nu+1}, \cdots, i_\mu) \\
&\quad\quad + k_\nu(i_1, i_2, \cdots, i_{\nu-1}, i_\nu, i_{\nu+1}, \cdots, i_\mu)\}]\cdot \\
&\quad\quad y(i_1, i_2, \cdots, i_\nu, \cdots, i_\mu) = 0
\end{aligned} \tag{7}$$

For brevity we let

$$\begin{aligned}
(i_1, i_2, \cdots, i_\mu) &= -s(i_1, i_2, \cdots, i_\mu) \\
&\quad - \textstyle\sum_{\nu=1}^{\mu} \{k_\nu(i_1, i_2, \cdots, i_{\nu-1}, i_\nu - 1, i_{\nu+1}, \cdots, i_\mu) \\
&\quad\quad + k_\nu(i_1, i_2, \cdots, i_{\nu-1}, i_\nu, i_{\nu+1}, \cdots, i_\mu)\}
\end{aligned} \tag{8}$$

It is to be observed that (7) is a $2\mu + 1$ point equation—that is y appears explicitly in the equation at $2\mu + 1$ points.

We shall consider (7) [(1)] subject to the boundary conditions

$$\begin{aligned}
y(0, i_2, \cdots, i_\mu) &= p_1\, y(1, i_2, \cdots, i_\mu) \\
y(m_1 + 1, i_2, \cdots, i_\mu) &= q_1\, y(m, i_2, \cdots, i_\mu) \\
y(i_1, 0, i_3, \cdots, i_\mu) &= p_2\, y(i_1, 1, i_3, \cdots, i_\mu) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
y(i_1, i_2, \cdots, i_{\mu-1}, 0) &= p_\mu\, y(i_1, i_2, \cdots, i_{\mu-1}, 1) \\
y(i_1, i_2, \cdots, i_{\mu-1}, i_{m_\mu+1}) &= q_\mu\, y(i_1, i_2, \cdots, i_{\mu-1}, m_\mu),
\end{aligned} \tag{9}$$

where $p_j \leqq 0, q_j \leqq 0, j = 1, 2, \cdots, \mu$.

We call the point set determined by (2) the FUNDAMENTAL DOMAIN for our problem.

Let us write down equation (7) for each of the points of the fundamental domain. We then substitute from boundary conditions (9) for each of the values

of y on the boundary of the fundamental domain. We have m homogeneous linear equations in m variables. A necessary and sufficient condition for the existence of a solution not identically zero is the vanishing of the determinant of the coefficients. This determinant is called the characteristic determinant for the problem. The characteristic determinant equated to zero is called the characteristic equation. We shall consider it as an equation in λ/n^2. Its roots are called the characteristic values for the problem. The order in which the equations are written down or the order of the variables in the equations is not important for our purposes. We shall, however, designate each equation and the resulting row in the characteristic determinant by the $(i_1, i_2, \cdots, i_\mu)$ which occurs in $(\lambda/n^2)y(i_1, i_2, \cdots i_\mu)$. We then write the variables in the equations in this same order and designate each column by the $(i_1, i_2, \cdots, i_\mu)$ of $y(i_1, i_2, \cdots, i_\mu)$ which occurs in that column. To illustrate what is meant we write down the characteristic determinant for $\mu = 2, m_1 = m_2 = 3, y(0, i_2) = y(4, i_2) = y(i_1, 0) = y(i_1, 4) = 0$. The column designation is at the top and the row designation at the side.

	(3, 3)	(2, 3)	(1, 3)	(1, 2)	(1, 1)	(3, 2)	(2, 2)	(3, 1)	(2, 1)
(3, 3)	$\frac{\lambda}{n^2} + r(3, 3)$	$k_1(2, 3)$	0	0	0	$k_2(3, 2)$	0	0	0
(2, 3)	$k_1(2, 3)$	$\frac{\lambda}{n^2} + r(2, 3)$	$k_1(1, 3)$	0	0	0	$k_2(2, 2)$	0	0
(1, 3)	0	$k_1(1, 3)$	$\frac{\lambda}{n^2} + r(1, 3)$	$k_2(1, 2)$	0	0	0	0	0
(1, 2)	0	0	$k_2(1, 2)$	$\frac{\lambda}{n^2} + r(1, 2)$	$k_2(1, 1)$	0	$k_1(1, 2)$	0	0
(1, 1)	0	0	0	$k_2(1, 1)$	$\frac{\lambda}{n^2} + r(1, 1)$	0	0	0	$k_1(1, 1)$
(3, 2)	$k_2(3, 2)$	0	0	0	0	$\frac{\lambda}{n^2} + r(3, 2)$	$k_1(2, 2)$	$k_2(3, 1)$	0
(2, 2)	0	$k_2(2, 2)$	0	$k_1(1, 2)$	0	$k_1(2, 2)$	$\frac{\lambda}{n^2} + r(2, 2)$	0	$k_2(2, 1)$
(3, 1)	0	0	0	0	0	$k_2(3, 1)$	0	$\frac{\lambda}{n^2} + r(3, 1)$	$k_1(2, 1)$
(2, 1)	0	0	0	0	$k_1(1, 1)$	0	$k_2(2, 1)$	$k_1(2, 1)$	$\frac{\lambda}{n^2} + r(2, 1)$

The characteristic determinant will be denoted by $\psi(\lambda/n^2)$.

Theorem: The characteristic determinant is symmetrical with reference to its principal diagonal.

To show this, write equation (7) for $(i_1, i_2, \cdots, i_{\nu-1}, i_\nu, i_{\nu+1}, \cdots, i_\mu)$ and then for $(i_1, i_2, \cdots, i_{\nu-1}, i_\nu + 1, i_{\nu+1}, \cdots, i_\mu)$. It is then immediate from the first of these equations that $k_\nu(i_1, i_2, \cdots i_\nu, \cdots, i_\mu)$ is in the $(i_1, i_2, \cdots, i_\nu, \cdots, i_\mu)$ row and the $(i_1, i_2, \cdots, i_\nu + 1, \cdots, i_\mu)$ column. Likewise from the second equation it is immediate that it is in the $(i_1, i_2, \cdots, i_\nu + 1, \cdots, i_\mu)$ row and the $(i_1, i_2, i_\nu, \cdots, i_\mu)$ column.

Theorem: The roots of the characteristic equation are all positive.

The proof given for this theorem when $\mu = 2$ is quite general (See page 97 of reference (2)).

We now write the characteristic determinant as a polynomial in λ/n^2 and equate to zero. We have

$$(10)\quad \psi(\lambda/n^2) = (\lambda/n^2)^m + A_1(\lambda/n^2)^{m-1} + A_2(\lambda/n^2)^{m-2} + \cdots + A_m = 0$$

Here A_j is the sum of the principal minors of order j taken from that determinant which is formed when we replace λ by zero in the characteristic determinant [(3) page 126)]. It is agreed that the sum of the principal minors of order zero is 1.

We now transform equation (10), multiplying its roots by $1/m$. We get

$$(11) \qquad (\lambda/n^2)^m + m^{-1}A_1(\lambda/n^2)^{m-1} + \cdots + m^{-m}A_m = 0$$

Now divide equation (11) through by $\left(\frac{\lambda}{n^2}\right)^m$ and denote $\frac{n^2}{\lambda}$ by ρ. We get

$$(12) \qquad 1 + m^{-1}A_1\rho + m^{-2}A_2\rho^2 + \cdots + m^{-m}A_m\rho^m = 0$$

We note particularly that the values of ρ which satisfy this equation are the reciprocals of the characteristic values after these characteristic values have been multiplied by $1/m$.

Our purpose is to consider equation (12) when n becomes infinite and to draw other conclusions relative to the characteristic values when n becomes infinite.

Let us consider A_1/m. Then

$$A_1 = \sum r(i_1, i_2, \cdots, i_\mu)$$

where the summation is extended over all points of the fundamental domain. Hence

$$\frac{1}{m}A_1 = \sum_{i_1=1}^{m_1}\sum_{i_2=1}^{m_2}\cdots\sum_{i_\mu=1}^{m_\mu} r(i_1, i_2, \cdots, i_\mu)\frac{1}{m_1}\cdot\frac{1}{m_2}\cdot\cdots\frac{1}{m^\mu}.$$

Hence

$$\lim_{n\to\infty}\frac{1}{m}A_1 = \frac{1}{M}\int_0^{M_1}\int_0^{M_2}\cdots\int_0^{M_{\mu-1}}\int_0^{M_\mu} R(\chi_1, \chi_2, \cdots, \chi_{\mu-1}, \chi_\mu)d\chi_\mu d\chi_{\mu-1}\cdots d\chi_2 d\chi_1$$

where R bears the same relation to r that K does to k.

We next consider $1/m^2A_2$. We know that A_2 is the sum of the principal minors of order 2. This is a sum of determinants of the following types:

(a) $$\begin{vmatrix} r(p_1, p_2, \cdots, p_\mu) & 0 \\ 0 & r(q_1, q_2, \cdots, q_\mu) \end{vmatrix}$$

where $|p_1 - q_1| > 1, \cdots, |p_\mu - q_\mu| > 1$.

(b) $$\begin{vmatrix} r(p_1, p_2, \cdots, p_\mu) & k_\nu(j_1, j_2, \cdots, j_\mu) \\ k_\nu(i_1, i_2, \cdots, i_\mu) & r(q_1, q_2, \cdots, q_\mu) \end{vmatrix}$$

where at least one of the following inequalities holds: $|p_1 - q_1| \leqq 2, \cdots, |p_\mu - q_\mu| \leqq 2$. Summation is to be taken over all possible selections of $(p_1, p_2, \cdots, p_\mu)$ and $(q_1, q_2, \cdots, q_\mu)$ of the fundamental domain. The sum of determinants of type (a) multiplied by $1/m^2 = 1/M^2 \cdot 1/n^{2\mu}$ approaches

$$\frac{1}{M^2}\int_0^{M_1}\int_0^{x_1^{(1)}}\int_0^{M_2}\int_0^{x_2^{(1)}}\cdots\int_0^{M_\mu}\int_0^{x_\mu^{(1)}} R(x_1^{(1)}, x_2^{(1)}, \cdots x_\mu^{(1)}).$$

$$R(x_1^{(2)}, x_2^{(2)}, \cdots, x_\mu^{(2)})dx_\mu^{(2)}dx_\mu^{(1)}\cdots dx_2^{(2)}dx_2^{(1)}dx_1^{(2)}dx_1^{(1)}$$

The sum of determinants of type (b) multiplied by $1/m^2$ approaches zero. This follows from the fact that they are bounded in absolute value and that the number of determinants of this type is of order n^μ while $1/m^2$ is of order $n^{2\mu}$.

Similarly $(1/m^3)A_3$ approaches

$$\frac{1}{M^3}\int_0^{M_1}\int_0^{x_1^{(1)}}\int_0^{x_1^{(2)}}\int_0^{M_2}\int_0^{x_2^{(1)}}\int_0^{x_2^{(2)}}\cdots\int_0^{M_\mu}\int_0^{x_\mu^{(1)}}\int_0^{x_\mu^{(2)}} R(x_1^{(1)}, x_2^{(1)}, \cdots x_\mu^{(1)}).$$

$$R(x_1^{(2)}, x_2^{(2)}, \cdots, x_\mu^{(2)})\cdot R(x_1^{(2)}, x_2^{(3)}, \cdots, x_\mu^{(3)})dx_\mu^{(3)}dx_\mu^{(2)}dx_\mu^{(1)}$$

$$\cdots dx_2^{(3)}dx_2^{(2)}dx_2^{(1)}x_1^{(3)}dx_1^{(2)}dx_1^{(1)}.$$

Proceeding in this way we have the following theorem

Theorem:

$$\text{(12)}\quad \lim_{n\to\infty}\frac{1}{m^\nu}A_\nu = \frac{1}{M^\nu}\int_0^{M_1}\int_0^{x_1^{(1)}}\cdots\int_0^{x_1^{(\nu)}}\cdots\int_0^{M_\mu}\int_0^{x_\mu^{(1)}}\int_0^{x_\mu^{(\nu)}}$$

$$R(x_1^{(1)}, x_2^{(1)}, \cdots, x_\mu^{(1)})\cdots R(x_1^{(\nu)}, x_2^{(\nu)}, \cdots, x_\mu^{(\nu)}).$$

$$dx_\mu^{(\nu)}\cdots dx_\mu^{(2)}dx_\mu^{(1)}\cdots dx_1^{(\nu)}\cdots dx_1^{(2)}dx_1^{(1)}.$$

Now let

$$B_\nu = \lim_{n\to\infty}\frac{1}{m^\nu}A_\nu$$

Theorem: The power series

$$1 + B_1\rho + B_2\rho^2 + \cdots + B_1\rho^\nu + \cdots$$

converges for all values of ρ

To prove this suppose

$$|R(x_1, x_2, \cdots, x_\mu)| < P$$

Then from (12)

$$B_\nu \leqq \frac{1}{M^\nu}\left(\frac{1}{\nu!}\right)^\mu P^\nu$$

from which the theorem follows.

Now let

$$\varphi(\rho) = 1 + B_1\rho + B_2\rho^2 + \cdots + B_\nu\rho^\nu + \cdots$$

and let

$$X(\rho, n) = 1 + \frac{1}{m}A_1\rho + \frac{1}{m^2}A_2\rho^2 + \cdots + \frac{1}{m^m}A_m\rho^m$$

Let Q be any positive number

Theorem:

$$\text{Limit}_{n\to\infty} X(\rho, n) = \varphi(\rho)$$

uniformly in ρ when

$$0 \leqq \rho \leqq Q$$

Proof: Consider $m^{-\nu}A_\nu\rho_\nu$. We know that A_ν consists of the sum of the principal minors of order ν of the determinant A_μ. The number of such principal minors is the number of combinations of m things ν at a time. Each such minor has at most $(2\mu + 1)$ non-zero elements in any row or column. Let the absolute value of each of these elements be less than J, then the absolute value of each of the principal minors is less than $(2\mu + 1)^\nu J^\nu$. Consequently

$$\left|\frac{1}{m^\nu} A_\nu \rho^\nu\right| < \frac{m(m-1)\cdots(m-\nu+1)}{\nu!}\left[\frac{(2\mu+1)J\rho}{m}\right]^\nu$$

$$< \frac{1}{\nu!}[(2\mu+1)J\rho]^\nu < \frac{1}{\nu!}[(2\mu+1)J\cdot Q]^\nu$$

This is the general term of a convergent series of positive constants. The theorem follows.

Theorem: The arithmetic mean of the characteristic values approaches a limit different from zero when n becomes infinite.

This theorem is just a restatement of the fact that $m^{-1}A_1$ approaches a limit different from zero. This limit is not zero since $R < 0$. We could state similar theorems from the facts that $m^{-\nu}A_\nu$ approach limits for other values of ν.

From the fact that $m^{-1}A_1$ approaches a non-zero limit and the fact that the characteristic values are all positive we have the following two theorems.

Theorem: Not all the characteristic values approach zero when n becomes infinite.

Theorem: Not all the characteristic values become infinite when n becomes infinite.

In case R is constant the integration indicated in (12) can be readily performed. For example if $\mu = 1$ then $B_\nu = R^\nu/\nu!$ and $\varphi(\rho) = e^{R\rho}$.

Now if we consider the equation $\Delta^2 y(i-1) + (\lambda/n^2)y(i) = 0$ subject to the boundary conditions

$$y(0) = y(n+1) = 0$$

the characteristic values are*

$$4\sin^2\frac{\pi j}{n+1}, \qquad j = 1, 2, \cdots, n$$

The arithmetic mean of these values is 2. Moreover $(1/n)\ 4\sin^2 \pi j/(n+1)$ approaches 0 for any j and $n/4\sin^2 \pi j/(n+1)$ becomes infinite for any j. This agrees with the fact that $e^{R\rho}$ is not zero for any ρ.

*(1), page 170.

REFERENCES

1. TOMLINSON FORT: *Finite Differences and Difference Equations in the Real Domain*, Clarendon 1948.
2. TOMLINSON FORT: *The Loaded Vibrating Net and resultingo Bundary-Value Problems for a Partial Difference Equation of the Second Order*, Journal of Math. and Physics, Vol. XXXIII, No. 1.
3. G. KOWALEWSKI: *Determinanten Theorie*, Leipsig 1909.

UNIVERSITY OF SOUTH CAROLINA
OFFICE OF ORDNANCE RESEARCH

(Received June 27, 1953)

IMPROVED APPROXIMATIONS TO DIFFERENTIAL EQUATIONS BY DIFFERENCE EQUATIONS[1]

By W. Robert Mann, C. L. Bradshaw and J. Grady Cox

It is well known that there are many difference equations which can be regarded as approximations to a given differential equation. By employing difference quotients of sufficiently high order one can, generally speaking, make the truncation error in the difference equation arbitrarily small. But unfortunately the use of higher order differences frequently brings one face to face with difficulties more serious than the obvious one resulting from the more complicated difference equation. For one thing, it is necessary to extemporize extra boundary conditions to determine the increased number of arbitrary constants in the general solution, and for another thing, the desired convergence-stability properties may be adversely affected or lost. The purpose of this paper is to illustrate a method for reducing truncation error in the difference equation without introducing differences of higher order, in any one index, than the order of the differential equation. For other work which has been done in this direction see [1], Chapters 8 and 10. It will be clear from the following presentation that the method is not equally adaptable to all problems, but numerical results are given which indicate that it promises significant improvement in some important problems where it can be used. The idea is simple and can best be presented by an example.

Consider the two-dimensional equation

$$L[\varphi] + \lambda^2\varphi = \varphi_{xx} + \varphi_{yy} + \lambda^2\varphi = 0. \tag{1}$$

Using the same mesh spacing, h, in both the x and y directions we get the usual finite difference approximation,

$$\Phi_{i+1,j} + \Phi_{i-1,j} + \Phi_{i,j+1} + \Phi_{i,j-1} - 4\Phi_{i,j} + \lambda^2h^2\Phi_{i,j} = M[\Phi] + \lambda^2h^2\Phi_{i,j} = 0. \tag{2}$$

Now let $\varphi(x, y)$ denote a solution of (1) corresponding to the characteristic number λ_0. Using the Taylor expansion to express $M[\varphi]$, we find that

$$M[\varphi] = -\lambda_0^2h^2\varphi_{i,j} + \frac{h^4}{12}\left(\frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4}\right)_{i,j}\varphi + \mathcal{O}(h^6).$$

In other words, the pair $[\varphi, \lambda_0]$ fails to satisfy (2) by an error term which is of order h^4. Referring back to (1) we find that

$$\varphi_{xxxx} + \varphi_{yyyy} = -2\varphi_{xxyy} - \lambda_0^2(\varphi_{xx} + \varphi_{yy}).$$

This leads to the following finite difference approximation to $\frac{1}{12}h^4(\varphi_{xxxx} + \varphi_{yyyy})$:

[1] This research was supported in part by the United States Air Force through the Office of Scientific Research of the Air Research and Development Command.

$$\frac{h^4}{12}(\varphi_{xxxx} + \varphi_{yyyy}) \simeq -\frac{1}{6}\Big[(\varphi_{i+1,j+1} - 2\varphi_{i,j+1} + \varphi_{i-1,j+1}) - 2(\varphi_{i+1,j} - 2\varphi_{i,j} + \varphi_{i,j-1}) + (\varphi_{i+1,j+1} - 2\varphi_{i,j-1} + \varphi_{i-1,j-1}) - \frac{\lambda_0^2}{12}(\varphi_{i+1,j} + \varphi_{i-1,j} + \varphi_{i,j+1} + \varphi_{i,j-1} - 4\varphi_{i,j})\Big].$$

Subtracting this expression in Φ and λ from the left-hand side of (2) and simplifying leads to the equation

$$(\tfrac{2}{3}\lambda^2h^2 - \tfrac{1}{3}\lambda_0^2)\Phi_{i,j} + (\tfrac{2}{3} + \tfrac{1}{12}\lambda^2h^2)(\Phi_{i-1,j} + \Phi_{i+1,j} + \Phi_{i,j+1} + \Phi_{i,j-1}) + \tfrac{1}{6}(\Phi_{i+1,j+1} + \Phi_{i+1,j-1} + \Phi_{i-1,j+1} + \Phi_{i-1,j-1}) = 0, \tag{3}$$

which we shall denote more briefly as follows:

$$(\tfrac{2}{3}\lambda^2h^2 - \tfrac{1}{3}\lambda_0^2)\Phi_{i,j} + (\tfrac{2}{3} + \tfrac{1}{12}\lambda^2h^2)(\Phi_{i\pm1,j} + \Phi_{i,j\pm1}) + \tfrac{1}{6}\Phi_{i\pm1,j\pm1} = 0. \tag{3}$$

If the left-hand side of (3) is denoted by $M[\Phi, \lambda]$, we find by substitution that

$$M[\varphi, \lambda_0] = E \cdot h^6 + \mathcal{O}(h^8),$$

where $E = -(3\lambda_0^4/6!)(\varphi_{xx} + \varphi_{yy}) - (14\lambda_0^2/6!)\varphi_{xxyy}$.

Equations (2) and (3) can both be regarded as finite difference approximations to (1). Both are of second order, but (3) is better in the sense that it has a truncation error of order h^6, whereas the error in the usual approximation (2) is of order h^4. The star of (2) consists of (i, j) together with its four closest neighbors, whereas the star of (3) includes also the vertices of the square about (i, j) as center. Although (3) represents a considerable improvement over (2), one can obtain a still better approximation to (1) without going beyond differences of the second order. An approximation to the dominant error term, Eh^6 in (3) is

$$E \simeq -\frac{3\lambda_0^4h^4}{6!}[\varphi_{i\pm1,j} + \varphi_{i,j\pm1} - 4\varphi_{i,j}] - \frac{14\lambda_0^2h^2}{6!}[\varphi_{i\pm1,j\pm1} - 2(\varphi_{i\pm1,j} + \varphi_{i,j\pm1}) + 4\varphi_{i,j}].$$

Subtracting this expression in Φ and λ from the left-hand side of (3) and simplifying gives the equation

$$\left(-\frac{10}{3} + \frac{67}{90}\lambda^2h^2 - \frac{12\lambda^4h^4}{6!}\right)\Phi_{i,j} + \left(\frac{2}{3} + \frac{2\lambda^2h^2}{45} + \frac{3\lambda^4h^4}{6!}\right)\cdot(\Phi_{i\pm1,j} + \Phi_{i,j\pm1}) + \left(\frac{1}{6} + \frac{14\lambda^2h^2}{6!}\right)\Phi_{i\pm1,j\pm1} = 0. \tag{4}$$

Denoting the left-hand side of (4) by $M_1[\Phi, \lambda]$ and substituting the pair $[\varphi, \lambda_0]$ we find that

$$M_1[\varphi, \lambda_0] = \mathcal{O}(h^8).$$

Equation (4) is obviously of second order and has the same star as (3), and yet the truncation error is only of order h^8.

We shall now apply the technique we have been describing to an artificially simple reactor problem chosen to facilitate comparison between the analytic solution of the differential boundary value problem and the numerical approximation to solutions of various associated finite difference problems. The numerical data were obtained with the ORACLE of the Oak Ridge National Laboratory. The test problem is the following:

$$\varphi_{xx} + \varphi_{yy} + \lambda^2\varphi = 0 \begin{cases} 0 < x < a \\ 0 < y < 1, \end{cases} \tag{5}$$

$$\varphi_{xx} + \varphi_{yy} - K^2\varphi = 0 \begin{cases} a < x < 1 \\ 0 < y < 1, \end{cases} \tag{6}$$

$$\varphi(0, y) = \varphi(1, y) = \varphi(x, 0) = \varphi(x, 0) = 0, \tag{7}$$

$$D_C\varphi'(a - 0, y) = D_R\varphi'(a + 0, y), \tag{8}$$

$$\varphi(a - 0, y) = \varphi(a + 0, y). \tag{9}$$

The quantities D_C, D_R and K are known, and we shall restrict our attention to the problem of finding the fundamental eigenvalue λ_0, i.e., the smallest admissible value of λ. The analytic approach shows that λ_0 can be determined from

$$\tan ax + Lx = 0, \tag{10}$$

where

$$x = \sqrt{\lambda_0^2 - \pi^2}, \quad \text{and} \quad L = \frac{D_1 \tanh \sqrt{K^2 + \pi^2}(1 - a)}{D_2\sqrt{K^2 + \pi^2}}. \tag{11}$$

The usual finite difference approximations to (5) and (6) are

$$\Phi_{i+1,j} + \Phi_{i-1,j} + \Phi_{i,j-1} + \Phi_{i,j+1} - 4\Phi_{i,j} + \lambda^2h^2\Phi_{i,j} = 0, \tag{12}$$

and

$$\Phi_{i+1,j} + \Phi_{i-1,j} + \Phi_{i,j+1} + \Phi_{i,j-1} - 4\Phi_{i,j} - K^2h^2\Phi_{i,j} = 0. \tag{13}$$

The normal derivative at the interface was approximated by

$$\varphi_x \simeq \frac{1}{2h}(\varphi_{i+1,j} - \varphi_{i-1,j}), \tag{14}$$

which leads to

$$D_C(\Phi^*_{M+1,j} - \Phi_{M-1,j}) = D_R(\Phi_{M+1,j} - \Phi^*_{M-1,j}), \qquad \text{where } Mh = a, \tag{15}$$

as an approximation to (8). The fictitious fluxes, $\Phi^*_{M+1,j}$ and $\Phi^*_{M-1,j}$, satisfy

$$\Phi^*_{M+1,j} + \Phi_{M-1,j} + \Phi_{M,j-1} + \Phi_{M,j+1} - 4\Phi_{M,j} + \lambda^2h^2\Phi_{M,j} = 0, \tag{16}$$

and

$$\Phi_{M+1,j} + \Phi^*_{M-1,j} + \Phi_{M,j+1} + \Phi_{M,j-1} - 4\Phi_{M,j} - K^2h^2\Phi_{M,j} = 0. \tag{17}$$

By eliminating the fictitious fluxes, we obtain

$$\frac{2D_2}{D_1+D_2}\Phi_{M+1,j}+\frac{2D_1}{D_1+D_2}\Phi_{M-1,j}+\Phi_{M,j\pm1}$$
$$-\left(4+\frac{D_2K^2h^2}{D_1+D_2}\right)\Phi_{M,j}+\frac{D_1\lambda^2h^2}{D_1+D_2}\Phi_{M,j}=0 \tag{18}$$

as the interface condition corresponding to (8). In place of (7) we have

$$\Phi_{0,j}=\Phi_{N,j}=\Phi_{i,0}=\Phi_{i,N}=0, \qquad \text{where } Nh=1. \tag{19}$$

The difference problem defined by the equations (12), (13), (18) and (19) is equivalent to a linear algebraic system of the form,

$$A\Phi+\lambda^2B\Phi=0, \tag{20}$$

where A is a square matrix whose order is the number of internal mesh points to be used, and B is a singular diagonal matrix of the same order as A. A fails to be symmetric because of the asymmetry in the relation which holds across the interface. An iterative procedure based on the following steps was used:

a) A vector Φ^0 is assumed.
b) The vector $D^{(0)}=\lambda^2B\Phi^0$ is computed, using a first guess $\lambda_0^{(0)}$ for the fundamental eigen number.
c) The system $A\Phi^{(1)}+D^{(0)}=0$ is solved.
d) The vector $D_*^{(1)}=B\Phi^{(1)}$ is computed.
e) The quantity $[\lambda_0^{(1)}]^2=\sum D^{(0)}/\sum D_*^{(1)}$ is found.
f) A new $D^{(1)}=[\lambda_0^{(1)}]^2D_*^{(1)}$ is found and the cycle is reentered at (c) and the process is continued till convergence to λ_0^2 is satisfactory.

Solution of the system $A\Phi+D=0$ was the major problem as far as computing time was concerned. The method used to solve this system was that of Richardson as refined by David Young [2]. Using this method and about 300 mesh points, the value of the fundamental eigen number obtained analytically could be duplicated to as many as 8 or 9 digits in some 15 minutes of computing time. After some practice in ordering the relaxation factors and progressively tightening the convergence criteria, about 300 iterations on the system $A\Phi+D=0$ were found to be sufficient.

The problem defined by equations (12), (13), (18) and (19) is based on difference approximations having a truncation error of order h^4. Referring back to (3) we see that

$$(4+\tfrac{1}{2}\lambda^2h^2)(\Phi_{i\pm1,j}+\Phi_{i,j\pm1})+(4\lambda^2h^2-20)\Phi_{i,j}+\Phi_{i\pm1,j\pm1}=0, \tag{21}$$

and

$$(4-\tfrac{1}{2}K^2h^2)(\Phi_{i\pm1,j}+\Phi_{i,j\pm1})-(4K^2h^2+20)\Phi_{i,j}+\Phi_{i\pm1,j\pm1}=0 \tag{22}$$

approximate to (5) and (6) with a truncation error which is only of order h^6. To obtain a refined approximation to the normal derivative at the interface we begin with the observation that

$$1/2h(\varphi_{i+1,j} - \varphi_{i-1,j}) = \varphi_x + \tfrac{1}{6}h^2\varphi_{xxx} + \mathcal{O}(h^4). \tag{23}$$

To the left of the interface we have that $\varphi_{xxx} = -\varphi_{yyx} - \lambda^2\varphi_x$.

$$\therefore \tfrac{1}{6}h^2\varphi_{xxx} \simeq -(1/12h)[\varphi_{i+1,j\pm1} - \varphi_{i-1,j\pm1} - 2(\varphi_{i+1,j} - \varphi_{i-1,j})] - \tfrac{1}{12}\lambda^2 h(\varphi_{i+1,j} - \varphi_{i-1,j}).$$

Subtracting this expression from the left-hand side of (23) we find that

$$\varphi_x = \left(\frac{1}{3h} + \frac{\lambda^2 h}{12}\right)(\varphi_{i+1,j} - \varphi_{i-1,j}) + \frac{1}{12h}(\varphi_{i+1,j\pm1} - \varphi_{i-1,j\pm1}) + \mathcal{O}(h^4). \tag{24}$$

We can now write the appropriate approximation to (8) to go with (21) and (22), namely

$$\begin{aligned} &D_C\left(\frac{1}{3h} + \frac{\lambda^2 h}{12}\right)(\Phi^*_{M+1,j} - \Phi_{M-1,j}) + \frac{D_C}{12h} \\ &\qquad \cdot (\Phi^*_{M+1,j+1} + \Phi^*_{M+1,j-1} - \Phi_{M-1,j+1} - \Phi_{M-1,j-1}) \\ &\qquad = D_R\left(\frac{1}{3h} - \frac{K^2 h}{12}\right)(\Phi_{M+1,j} - \Phi^*_{M-1,j}) \\ &\qquad + \frac{D_R}{12h}(\Phi_{M+1,j+1} + \Phi_{M+1,j-1} - \Phi^*_{M-1,j+1} - \Phi^*_{M-1,j-1}). \end{aligned} \tag{25}$$

The difference problem defined by (21), (22), (25) and (19) can be solved exactly. Let us call the region to the left of the interface the core, and the region to the right, the reflector. Let ${}_C\Phi_{i,j}$ denote the flux in the core and ${}_R\Phi_{i,j}$, that in the reflector. Then

$${}_C\Phi_{i,j} = C_C \sin \beta hi \sin \pi hj \tag{26}$$

and

$${}_R\Phi_{i,j} = C_R \sinh (N - i)\gamma h \sin \pi hj \tag{27}$$

where

$$\gamma = \frac{1}{h} \operatorname{arc\,cosh}\left[\frac{20 + 4K^2h^2 + (K^2h^2 - 8)\cos \pi h}{8 - K^2h^2 + 4 \cos \pi h}\right] \tag{28}$$

and

$$\beta = \frac{1}{h} \arccos\left[\frac{20 - 8 \cos \pi h + \lambda_0^2 h^2(1 - \cos \pi h) - 5\lambda_0^2 h^2}{8 + \lambda_0^2 h^2 + 4 \cos \pi h}\right] \tag{29}$$

Instead of eliminating the fictitious fluxes from (25) we determine them using (26) and (27) as follows:

$$\Phi^*_{M+1,p} = C_C \sin \beta h(M + 1) \sin \pi hp$$

$$\Phi^*_{M-1,p} = C_R \sinh (N - M + 1)\gamma h \sin \pi hp$$

TABLE I

D_C/D_R	Exact λ_0^2/K^2	$1/h$	Crude—Crude λ_0^2/K^2	Error	Refined—Crude λ_0^2/K^2	Error	Refined—Refined λ_0^2/K^2	Error
5	22.3852	10	22.2541	−.1311	22.4839	+.0987	22.3876	+.0024
		18	22.3447	.0405	22.4150	.0298	22.3854	.0002
		20	22.3524	.0328	22.4093	.0241	22.3853	.0001
		30	22.3706	.0146	22.3959	.0107	22.3852	.0000
		40	22.3770	.0082	22.3912	.0060	22.3852	.0000
		50	22.3799	−.0053	22.3890	+.0038	22.3852	.0000
4	22.9898	10	22.8654	−.1244	23.1120	+.1222	22.9928	+.0030
		18	22.9514	.0384	23.0268	.0370	22.9901	.0003
		20	22.9587	.0311	23.0197	.0299	22.9900	.0002
		30	22.9760	.0138	23.0031	.0133	22.9899	.0001
		40	22.9820	.0078	22.9973	.0075	22.9898	.0000
		50	22.9848	−.0050	22.9946	+.0048	22.9898	.0000
3	23.9513	10	23.8368	−.1145	24.1111	+.1598	23.9551	+.0038
		18	23.9160	.0353	23.9997	.0484	23.9517	.0004
		20	23.9227	.0286	23.9904	.0391	23.9516	.0003
		30	23.9386	.0145	23.9686	.0173	23.9514	.0001
		40	23.9442	.0071	23.9611	.0098	23.9513	.0000
		50	23.9468	−.0045	23.9576	+.0063	23.9513	.0000
2	25.7143	10	25.6146	−.0997	25.9429	+.2286	25.7199	+.0056
		18	25.6836	.0307	25.7835	.0692	25.7149	.0006
		20	25.6895	.0248	25.7703	.0560	25.7147	.0004
		30	25.7033	.0110	25.7391	.0248	25.7144	.0001
		40	25.7081	.0062	25.7282	.0139	25.7143	.0000
		50	25.7104	−.0039	25.7232	+.0089	25.7143	.0000
1	29.9536	10	29.8591	−.0945	30.3329	+.3793	29.9637	+.0101
		18	29.9249	.0486	30.0686	.1150	29.9546	.0010
		20	29.9304	.0232	30.0466	.0930	29.9542	.0006
		30	29.9433	.0103	29.9948	.0412	29.9537	.0001
		40	29.9478	.0058	29.9767	.0231	29.9536	.0000
		50	29.9499	−.0037	29.9684	+.0148	29.9536	.0000
$\frac{1}{2}$	35.3444	10	35.1508	−.1936	35.8377	+.4933	35.3591	+.0147
		18	35.2864	.0580	35.4952	.1508	35.3459	.0015
		20	35.2975	.0469	35.4665	.1221	35.3454	.0010
		30	35.3237	.0207	35.3986	.0542	35.3446	.0002
		40	35.3328	.0116	35.3749	.0305	35.3445	.0001
		50	35.3370	−.0074	35.3639	+.0195	35.3444	.0000
$\frac{1}{4}$	38.5124	10	38.1811	−.3313	39.0053	+.4929	38.5274	+.0150
		18	38.4127	.0997	38.6644	.1520	38.5139	.0015
		20	38.4318	.0806	38.6355	.1231	38.5134	.0010
		30	38.4768	.0356	38.5671	.0547	38.5126	.0002
		40	38.4924	.0200	38.5432	.0308	38.5125	.0001
		50	38.4996	−.0128	38.5321	+.0197	38.5124	.0000

TABLE I—*Continued*

D_C/D_R	Exact λ_0^2/K^2	$1/h$	Crude—Crude λ_0/K^2	Error	Refined—Crude λ_0^2/K^2	Error	Refined—Refined λ_0^2/K^2	Error
$\frac{1}{4}$	40.5540	10	40.0963	−.4577	41.0132	+.4592	40.5677	+.0137
		18	40.4155	.1385	40.6966	.1426	40.5553	.0013
		20	40.4420	.1120	40.6695	.1155	40.5549	.0009
		30	40.5044	.0496	40.6054	.0514	40.5542	.0002
		40	40.5261	.0279	40.5829	.0289	40.5541	.0001
		50	40.5362	−.0178	40.5725	+.0185	40.5540	.0000
$\frac{1}{2}$	41.9654	10	41.4021	−.5633	42.3848	+.4194	41.9773	+.0119
		18	41.7943	.1711	42.0964	.1310	41.9666	.0012
		20	41.8270	.1384	42.0716	.1062	41.9662	.0008
		30	41.9041	.0613	42.0127	.0473	41.9655	.0001
		40	41.9309	.0345	41.9920	.0266	41.9654	.0000
		50	41.9434	−.0220	41.9824	+.0170	41.9654	.0000
$\frac{1}{4}$	44.3878	10	43.6074	−.7804	44.7062	+.3184	44.3945	+.0067
		18	44.1489	.2389	44.4886	.1008	44.3884	.0006
		20	44.1944	.1934	44.4696	.0818	44.3882	.0004
		30	44.3020	.0858	44.4243	.0365	44.3878	.0000
		40	44.3395	.0483	44.4084	.0170	44.3878	.0000
		50	44.3569	−.0309	44.4009	+.0131	44.3878	.0000

Solution of the problem (19), (21), (22) and (25) for λ_0 gives

$$\tan a\beta + L\left[\frac{A_1 + \lambda_0^2 h^2}{A_1 - K^2h^2}\right]\sin(hx) = 0$$

where

$$L = \frac{D_C \tanh \gamma(1-a)}{D_R \sinh \gamma h}, \tag{30}$$

and

$$\lambda_0^2 = \frac{20 - 4\cos(\beta h)\cos \pi h - 8\cos(\beta h) - 8\cos \pi h}{h^2[\cos(\beta h) + \cos \pi h + 4]}.$$

On comparing (21) and (22) with (12) and (13) we see that the refined approximation at interior points is not very much more complicated than the crude approximation. The contrast between (15) and (25), however, shows that the higher order approximation at the interface points is achieved at the cost of a considerable increase in complexity. This observation suggests that we investigate the possibility of using (21) and (22) in conjunction with (15) instead of (25). Since interface points are not very numerous in comparison with interior points, it seemed reasonable to hope that this scheme might yield much greater accuracy at very small cost. The problem defined by (15), (19), (21) and (22) can be solved to give

$$\tan a\beta + L \sin h\beta = 0 \tag{32}$$

where β and L are given by (29) and (30), and λ_0 is given by (31).

We now proceed to the presentation of some relevant data in Table I. In order to facilitate a comparison of the various approximations discussed, we shall find it convenient to introduce some new terminology. The approximating problem defined by (12), (13), (18) and (19) will be called the crude–crude scheme to indicate that the truncation error in the equations is of order h^4 at both interior points and interface points. The approximating problem defined by (19), (21), (22) and (25) will be called the refined–refined scheme, and equations (15), (19), (21) and (22) will be said to define the refined–crude scheme. The names of these three schemes will be abbreviated further to (C–C), (R–R) and (R–C), respectively. Since the truncation error in the (R–R) scheme is everywhere only of order h^6 as compared with a truncation error of order h^4 in the (C–C) scheme, it seemed reasonable to hope that one could obtain about as much accuracy using (R–R) with $\sqrt{N}$ mesh points as by using (C–C) with N points. The data substantiate this hope to some extent. It is very clear, however, that the (R–C) scheme is a complete failure, i.e., there seems to be no advantage in reducing truncation error at interior points without achieving the same reduction at interface points. No data have been obtained using the scheme corresponding to (4) which has truncation error reduced to order h^8, and which seems to be the best possible approximation to (1) using second order differences. It is hoped that further exploration of this method of reducing truncation error will make feasible the solution of problems which would require working with prohibitively large algebraic systems if the usual crude difference approximations were used.

REFERENCES

1. W. E. Milne, *Numerical Solution of Differential Equations* (New York: John Wiley and Sons, Inc., 1953).
2. David Young, "On Richardson's Method for Solving Systems with Positive Definite Matrices," *Journal of Mathematics and Physics* **32** (1953–54), pp. 243–255.

University of North Carolina
Oak Ridge National Laboratory
Elgin Air Force Base

(Received February 21, 1956)